Protein Heteronuclear NMR Assignments Using Mean-Field Simulated Annealing

NICOLAS E. G. BUCHLER,* ERIK R. P. ZUIDERWEG,*,† HONG WANG,* AND RICHARD A. GOLDSTEIN*,‡,§

*Biophysics Research Division, †Department of Biological Chemistry, and ‡Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109-1055

Received July 29, 1996; revised December 23, 1996

A computational method for the assignment of the NMR spectra of larger (21 kDa) proteins using a set of six of the most sensitive heteronuclear multidimensional nuclear magnetic resonance experiments is described. Connectivity data obtained from HNC$_\alpha$, HN(CO)C$_\alpha$, HN(C$_\alpha$)H$_\alpha$, and H$_\alpha$(C$_\alpha$CO)NH and spin-system identification data obtained from CP-(H)CCH–TOCSY and CP-(H)C(C$_\alpha$CO)NH–TOCSY were used to perform sequence-specific assignments using a mean-field formalism and simulated annealing. This mean-field method reports the resonance assignments in a probabilistic fashion, displaying the certainty of assignments in an unambiguous and quantitative manner. This technique was applied to the NMR data of the 172-residue peptide-binding domain of the *E. coli* heat-shock protein, DnaK. The method is demonstrated to be robust to significant amounts of missing, spurious, noisy, extraneous, and erroneous data. © 1997 Academic Press

INTRODUCTION

The use of nuclear magnetic resonance spectroscopy to investigate the structure and dynamics of proteins generally starts with the determination of the correct assignment of the observed resonances to the individual nuclei in the protein, called a ‘‘sequence-specific assignment’’ (*1*). While the assignment procedure for smaller, unlabeled proteins has largely remained a manual task due to the intrinsic incompleteness of the homonuclear NMR spectra, the assignment for larger, labeled proteins is paradoxically more amenable to computer assistance and automation because of the spectral simplicity of the triple-resonance NMR experiments. Thus, quite a number of automated and semi-automated methods for performing sequence-specific assignments of labeled protein spectra have been described in recent years; these include the use of neural networks, genetic algorithms, simulated annealing, pseudo-energy minimization, and constraint satisfaction (*2*).

Most of these automated assignment methods were developed for proteins in the range 8–15 kDa and utilize a large number of source NMR spectra. For a variety of reasons, even modest increases in protein size greatly complicate the assignment process. First, the larger number of more poorly resolved resonances results in greater problems with spectral overlap. This situation is exacerbated by increased relaxation rates, which limit the source data to only those obtainable by the most sensitive experiments. Because of these limitations, any practical automated assignment method for proteins above 15 kDa must be able to work with data that are limited, probabilistic, ambiguous, and sometimes missing or inaccurate. A further problem is the existence of an exponentially large number of ways of assigning N spin systems to N residues in the protein, which eliminates any hope of performing an exhaustive search over all possible assignments, even using the fastest computers. In order to deal with this problem, many automated methods rely on some form of buildup procedure where the most certain sections of the protein are assigned first, consequently reducing the number of remaining possibilities. As even the more constrained regions cannot be definitively assigned before a more complete solution is found, however, these methods generally must include some way of going back and correcting errors, using techniques such as simulated annealing or genetic algorithms (*3, 4*). While such backtracking methods work well for smaller proteins, the number of possibilities that need to be considered again grows exponentially with the size of the system, making applications to larger proteins problematic. The difficulty is increased by the existence of suboptimal assignments that cannot be transformed into the correct assignment without passing through highly nonoptimal intermediate states.

We describe here the use of a mean-field approach to simulated annealing that is sensitive enough to yield good quality assignments for proteins in the 20 kDa range. In contrast to previous approaches, this method allows the entire assignment to evolve in a holistic manner, removing the need for proof-reading mechanisms. The mean-field approach is defined by considering the assignment *probabilistically,* allowing us to smooth the assignment space and lessen the multiple-minima problem. We first develop the

§ To whom correspondence should be addressed.

1090-7807/97 $25.00

methodological basis for the work, and then discuss its performance for 172 residues from a domain of the chaperone protein DnaK (21 kDa, $\tau_R = 15$ ns), restricting our source data to those derivable from the most sensitive triple-resonance experiments. We also demonstrate how the method is robust to uncertainties, ambiguities, and errors by using progressively degraded source data. The computer program can be obtained directly from the authors.

THEORY

Automated assignment procedures generally use through-bond connectivity data obtained from 2-, 3-, or 4-D triple-resonance experiments and are essentially based on the construction and linkage of NH-based ''T'' units. T_{C_α} units are constructed from HNC_α and $HN(CO)C_\alpha$ data and connect $C_\alpha(i-1)$, $N(i)$, $H(i)$, $C_\alpha(i)$; T_{H_α} units are constructed from $HN(C_\alpha)H_\alpha$ and $H_\alpha(C_\alpha CO)NH$ data and connect $H_\alpha(i-1)$, $N(i)$, $H(i)$, $H_\alpha(i)$. Further, T units such as T_{C_β} can be constructed from $HN(C_\alpha)C_\beta$ and $HN(COC_\alpha)C_\beta$ data to connect $C_\beta(i-1)$, $N(i)$, $H(i)$, $C_\beta(i)$, etc. It is the task of the automated assignment program to correctly link these different T units into strings such as $\cdots C_\alpha(i-1)$, $N(i)$, $H(i)$, $C_\alpha(i)$, $N(i+1)$, $H(i+1)$, $C_\alpha(i+1)$, $\cdots$ and to place these in the protein sequence.

To overcome the degeneracy of the chemical shifts at the ends of the individual T's, several different T types are generally linked simultaneously. Most reported automated assignment programs thus use at least three different T types: T_{C_α}, T_{H_α}, and T_{C_β} by the programs developed in the labs of Montelione and Müller (*5, 6*); T_{C_α}, T_{H_α}, and T_{CO} in the lab of Marion (*7*); and T_{C_α}, T_{CO}, and ^{15}N-TOCSY in the lab of Markley (*8*). Whenever T_{C_α}, T_{H_α}, and T_{C_β} are used, placement of the strings in the sequence is obtained implicitly via characteristic C_α/C_β shifts. The placement of these T's can also be based on C_α chemical shifts (*7*), or more generally, on side-chain shift signatures obtained from $(H)C(C_\alpha CO)NH$-type experiments (*9*). Table 1 shows computations of the first-scan sensitivities of some of the more commonly used experiments. Thus, sets of experiments that yield complete T_{C_α}, T_{H_α}, T_{CO}, or T_{C_β} units decrease in sensitivity, in that order. It follows, for instance, that the AUTOASSIGN program of Montelione and co-workers (*5*) can likely only be used for smaller proteins because it relies heavily on T_{C_β} units (called ''C_α ladders''), requiring $HN(C_\alpha)C_\beta/(H_\beta)C_\beta(C_\alpha)NH$ experiments which dramatically lose sensitivity when applied to larger molecules. Procedures avoiding the T_{C_β} and/or T_{CO} experiments should be chosen for larger proteins [unless C_α is deuterated (*11*)]. The set of T_{C_α} and T_{H_α} experiments used by the Fesik group certainly qualifies as a good choice for the analysis of the spectra of larger proteins; their (in principle desirable) use of two 4-D experiments in lieu of four 3-D experiments, however, limits intrinsic sensitivity considerably (*9*).

TABLE 1

First-Scan Sensitivities of Some of the NMR Experiments Most Commonly Used for Establishing Sequence-Specific Assignments, Relative to Single Proton Pulses, at Four Rotational Correlation Times Roughly Corresponding to Molecular Weights of 7, 13, 20 and 26 kDa

	τ_R (ns)			
Experiment	5 ns	10 ns	15 ns	20 ns
	T_{C_α} experiments			
$HN(CO)C_\alpha$	**0.45**	**0.32**	**0.17**	**0.11**
HNC_α	**0.40**	**0.28**	**0.13**	**0.09**
$(H_\alpha)C_\alpha(CO)HN$	0.45	0.27	0.14	0.08
$(H_\alpha)C_\alpha NH$	0.25	0.12	0.05	0.03
	T_{H_α} experiments			
$H_\alpha(C_\alpha CO)NH$	**0.45**	**0.27**	**0.14**	**0.08**
$HN(C_\alpha)H_\alpha$	**0.34**	**0.21**	**0.09**	**0.05**
$HN(COC_\alpha)H_\alpha$	0.38	0.23	0.11	0.06
$H_\alpha C_\alpha NH$	0.25	0.12	0.05	0.03
	T_{CO} experiments			
HNCO	0.49	0.37	0.22	0.15
$H_\alpha C_\alpha CO$	0.54	0.32	0.19	0.12
$HN(C_\alpha)CO$	0.26	0.14	0.05	0.02
	T_{C_β} experiments			
$(H_\beta)C_\beta(C_\alpha CO)NH$	**0.37**	**0.19**	**0.08**	**0.04**
$(H_\beta)C_\beta(C_\alpha)NH$	0.20	0.08	0.03	0.02
$HN(COC_\alpha)C_\beta$	0.25	0.12	0.04	0.02
$HN(C_\alpha)C_\beta$	0.22	0.10	0.03	0.01
	Other experiments			
(H)CCH	**0.43**	**0.23**	**0.12**	**0.06**
$H_\alpha C_\alpha(CO)N$	0.40	0.19	0.09	0.05
$H_\alpha C_\alpha N$	0.16	0.05	0.02	0.01

Note. Optimal concatenation of coherence-transfer periods (all HSQC-type) was assumed, and transfer times were optimized for every step using transverse relaxation rates at taken from (*10*). Linewidths of 2, 3, 4, and 6 Hz were assumed for the ^{13}CO coherence at the four correlation times, for 500 MHz equipment. Degradation of efficiency caused by the cumulative effects of RF inhomogeneity and resonance offsets in multiple-pulse experiments was not taken into account. As can be seen, the sensitivity of different experiments shows varying degrees of dependence on protein size. The experiments used for this study are indicated in boldface.

It should be stressed that, for larger proteins, neither spin system nor sequential information by itself is adequate to provide a unique sequence-specific assignment. Specifically, there are multiple locations in the sequence consistent with the spin-system identification data, and likewise many more possible sets of adjacent residues consistent with the connectivity data than actually exist in the protein. Selecting the assignment from the multiple possibilities that best fulfills *all* of these different sources of information forms the crux of the assignment problem.

We are generally interested in finding the ''best'' assignment according to some objective criterion. Often optimization problems can be reposed as energetic problems, allowing the development of thermodynamic analogies. For instance, a cost function can be defined as an ''energy,'' so that the search for an optimal solution can be rephrased as a search for a global energy minimum. Such approaches are especially useful for probabilistic information, as Boltzmann's equation can be inverted so that the energy corresponding to a given possibility is equal to the negative logarithm of the respective probability, in units of kT. (In practice, both k and T can be set equal to one.) In the current case, there are two different types of terms, representing the sequential connectivity and the side-chain identity. Sequential connectivity information arising from linking T_{C_α} and T_{H_α} is encoded as $E_{\mathrm{adj}}(i, i')$, the compatibility for spin system i to be followed by spin system i' in the sequence. Side-chain identity data can be translated into an energy parameter $E_{\mathrm{seq}}(i, j)$, which encodes the compatibility of spin system i to the residue types at sequence positions j and $j - 1$.

The complete sequence-specific assignment is a one-to-one mapping between spin systems and residues in the protein. We can define the assignment by specifying $\{L_i\}$, where L_i is the ''location'' of spin system i in the sequence. The total energy E as a function of the assignment can be expressed by summing over all possible assignments for all resonances while eliminating most of the terms with Kronecker deltas, which restricts the sums to those discrete terms representing where the residues are actually assigned:

$$E(\{L_i\}) = \sum_{i,j} E_{\mathrm{seq}}(i,j)\delta_{L_{i,j}} + \sum_{i,i',j} E_{\mathrm{adj}}(i,i')\delta_{L_{i,j}}\delta_{L_{i',j+1}}. \quad [1]$$

The assignment problem is the search for the energy minimum of this function in the multidimensional space of all possible assignments $\{L_i\}$. While the side-chain identification term $E_{\mathrm{seq}}(i, j)$ is only a function of the local assignment of spin set i to position j, the connectivity term $E_{\mathrm{adj}}(i, i')$ is nonlocal, in that it is a function of where two different spin systems are assigned. In addition, there will be nonlocality induced by the one-to-one nature of the mapping represented by $\{L_i\}$. As a result, the assignment of each spin system must depend on how all of the other spin systems are assigned. It is this aspect that is responsible for the multiple-minima nature of the energy function. The search for the optimal assignment is also greatly complicated by the limited move set for transforming one assignment into another. Assignments can generally only be changed by swapping sets of spin systems.

The heart of our new approach is in the use of a mean-field formalism for addressing the minimization problem. Rather than consider explicit assignments, as represented by $\{L_i\}$ above, we consider instead a large ensemble of assignments, as is done with approaches that use genetic algorithms (*4*). In contrast to genetic algorithms, however, we use the mean-field approximation and consider that each spin system in any assignment feels the *average* influence of all of the other assignments in the ensemble. We can then represent the whole ensemble as a single continuous assignment $\{P(i, j)\}$, where $P(i, j)$ represents the *occupancy* or the fraction of assignments where spin system i is assigned to sequence location j. The power of this approach lies in the fact that these occupancies are not restricted to the integer set $\{0, 1\}$ of the Kronecker delta $\delta_{L_{i,j}}$. By allowing the occupancies to vary continuously from zero to one, we are able to avoid the problems caused by a restricted set of possible moves caused by the discrete nature of the one-to-one mapping, allowing us to ''tunnel'' from one assignment to another through unphysical intermediates that do not correspond to possible assignments.

With this continuous assignment space, the Kronecker deltas in Eq. [1] are replaced by the assignment occupancies, yielding

$$E(\{P(i,j)\}) = \sum_{i,j} E_{\mathrm{seq}}(i,j)P(i,j) + \sum_{i,i',j} E_{\mathrm{adj}}(i,i')P(i,j)P(i',j+1). \quad [2]$$

As each spin system corresponds to one residue, and likewise, each residue corresponds to one spin system, for a protein of length N, $P(i, j)$ must satisfy $2N$ normalization constraints

$$\sum_i P(i,j) = 1 \quad (\forall j) \quad [3]$$

$$\sum_j P(i,j) = 1 \quad (\forall i). \quad [4]$$

The occupancies are the dynamic variables for the simulated annealing, and the areas of confident assignments should eventually converge to zero or one. The occupancies are initially defined by a uniform distribution (i.e., each spin system is distributed uniformly over all residues) and given ''velocities'' $\partial P(i, j)/\partial t$ drawn from a Maxwell–Boltzmann distribution at some temperature T. The ''forces'' acting on our dynamic occupancies are found by computing the negative gradient of the total energy function, where all ''masses'' have been set to one. The occupancies and velocities are updated by integration of Newton's equations of motion through use of the Verlet algorithm (*12*). Periodically, the velocities are rerandomized, as the temperature is

TABLE 2
Spin-System Classes Used for Identification of Residue Type from (H)CCH Data (i) and (H)C(C_αCO)NH Data ($i - 1$)

		Best		Good		Fair	
Class	Theoretical	i	$i - 1$	i	$i - 1$	i	$i - 1$
{A}	13	13	12	13	12	11	11
{T}	13	13	12	13	12	11	11
{G}	12	12	13	12	13	11	11
{S}	11	11	11	11	11	9	9
{P}	5	5		5			
{IPT}			5		5		
{ILV}	39	38	35	38	35	32	29
{QEM}	28	28	26				
{RK}	20	20	20				
{NDCQEHMFWY}	31	30	29				
{NDCQEHMFWYRK}				78	75	61	60
{~G, ~P}		2	2	2	2		
Unknown			7		7	37	41

Note. The second column gives the number of occurrences in the 172 amino acid fragment of DnaK. The following columns show the number of identifications in the ‘‘Best,’’ ‘‘Good,’’ and ‘‘Fair’’ quality of spectra.

gradually decreased to zero. The $2N$ holonomic constraints represented by Eqs. [3] and [4] and the N^2 nonholonomic constraints [$P(i, j) \geqslant 0$] are satisfied using an iterative Lagrange-multiplier approach.

RESULTS

The mean-field assignment program was demonstrated on the 172-residue fragment of the peptide-binding domain of the *E. coli* protein DnaK (21 kDa). DnaK is a heat-shock protein, a member of the Hsp-70 family, that has been shown to function as a chaperone for protein folding *in vivo.* The chemical-shift dispersion in DnaK is relatively poor, a situation exacerbated by large linewidths. Input peak lists were constructed directly from experimental data. Spin-system identification was obtained manually from a cross-polarization (CP)-driven (H)CCH experiment that correlated the resonances of aliphatic side-chain protons with backbone α-protons and α-carbons, and a CP-driven (H)C(C_αCO)NH experiment that correlated the resonances of side-chain carbons of residues $i - 1$ with backbone NHs of residues i (*13–15*). Many spin systems of the parent residue as well as that of the preceding residue could be uniquely identified, whereas others could be grouped together in different classes (see Table 2). In the ‘‘best’’ set, A, G, S, and T are uniquely identified, other methyl-containing residues are combined into the class {ILV}, the classic AMX systems into a single large class, K and R into a class {KR}, and the classic AMPX spin systems into a class {QEM}. Various computations with degraded versions of these data were also carried out, as described below. For each of these spin system identification sets, $E_{\text{seq}}(i, j)$ was set equal to some prohibitive value (+10) if the spin system was not compatible with the residue class at location j, and a large negative value (−4) if the system and residue class were compatible. Similarly, either +10 or −4 was added to $E_{\text{seq}}(i, j)$ if the information about the side-chain identity of the residue preceding spin system i in the sequence is consistent with the residue class at position $j - 1$. If a spin system was compatible with more than one class, −4 was added for each possibility. It would be possible to include more refined forms of this information encompassing varying degrees of confidence, by allowing a range of energy parameter values. Such confidence degrees can in principle be computed using Bayesian algorithms from chemical-shift data bases (*2*).

Sequential connections were established with a combination of the four most sensitive triple-resonance experiments that yield complete T sets: HNC_α, HN(CO)C_α, HN(C_α)H_α, and H_α(C_αCO)NH [(*15–20*); see Table 1]. This information was combined to provide T_{C_α} and T_{H_α} units consisting of the chemical shifts corresponding to the $C_\alpha(i)$, $C_\alpha(i - 1)$, $H_\alpha(i)$, and $H_\alpha(i - 1)$ spins respectively. The program used an unedited peak-pick table of these T_{C_α} and T_{H_α} frequencies. Possible adjacencies were constructed by observing that if spin system i is followed by spin system i' in the sequence, the $C_\alpha(i)$ and $H_\alpha(i)$ resonances must be the same within experimental resolution as the $C_\alpha(i' - 1)$ and $H_\alpha(i' - 1)$ resonances. Assuming that there is a Gaussian distribution of experimental measurements of the resonance position with widths σ_C and σ_H for the C_α and H_α resonances, respectively, then inversion of the Boltzmann equation results in an energy term quadratic in the difference in values for the resonance positions:

$$E_{\mathrm{adj}}(i, i') = \frac{1}{2\sqrt{2}\sigma_{\mathrm{H}}^{2}}[\mathrm{H}_{\alpha}(i) - \mathrm{H}_{\alpha}(i'-1)]^{2} + \frac{1}{2\sqrt{2}\sigma_{\mathrm{C}}^{2}}[\mathrm{C}_{\alpha}(i) - \mathrm{C}_{\alpha}(i'-1)]^{2}. \quad [5]$$

Prolines were specifically omitted from the sequence, because they have no sequential connectivity information in the four chosen triple-resonance experiments. Consequently, $E_{\mathrm{adj}}(i, i')P(i, j)P(i', j+1)$ is not included in the sum in Eq. [2] if a proline was excised between locations j and $j + 1$. The quality of the assignment was measured by q, defined as the average probability for the correct assignment

$$q = \frac{1}{N}\sum_{i} P[i, L_{c}(i)], \quad [6]$$

where $L_c(i)$ is the residue location that corresponds to spin system i when correctly assigned. The ''correct'' assignment was obtained previously from the same data using a semi-automated process using computer-graphic-driven interfaces (*15*), and was validated beyond any reasonable doubt from sequential NOE connectivities in ^{15}N- and ^{13}C-resolved NOESY data, compatibility with canonical secondary structure elements, and the structural analogy of these derived secondary structure elements with those of the homologous protein Hsc-70 (*21*). In repeated runs, the simulated annealing consistently found the correct assignment of the 172-residue DnaK protein based on the complete NMR data.

One common difficulty with NMR peak assignments is that the data may be ambiguous, noisy, incomplete, or erroneous. It is thus important for any computational method to be robust to these types of errors. In order to investigate the robustness of the mean-field simulated annealing approach, the data set was degraded in combinations of five ways:

Missing data. Either 10, 20, or 30% of the spin systems was deleted from the data set, and replaced with a set with no information [no observed peaks, hence no contribution from $E_{\mathrm{adj}}(i, i')$, and no preferences for any type of side chain in the protein, so E_{seq} was set equal to -4 for all locations in the sequence].

Extraneous data. Resonances corresponding to 30 or 45 spin systems picked from nonadjacent locations of a completely different protein [Human Stromelysin-1 (*20*)] were mixed in with the spin systems corresponding to DnaK. The extraneous data had both spectral and sequential identity.

Gaussian noise. Normally distributed noise was added to the chemical shifts of all of the protons (±0.03 ppm) and carbons (±0.10 ppm) used in the calculation of $E_{\mathrm{adj}}(i, i')$ (Eq. [5]).

Erroneous identification data. The identification data for six of the spin systems were modified to make these spin systems *incompatible* with their correct location.

TABLE 3
***q* Values for Runs with Various Degrees of Degradation**

Sidechain information	Other degredation	% of data missing 0	10	20	30
Best	None	1.00	1.00	0.99	0.86
Good	None	1.00	1.00	0.86	0.65
Fair	None	1.00	0.93	0.79	0.49
Best	Gaussian noise	1.00	0.99	0.99	0.92
Best	Erroneous identification data	1.00	1.00	0.90	0.92
	30 extra spin systems				
Best	None	1.00	0.99	0.95	0.78
Good	None	1.00	0.96	0.66	0.54
Fair	None	1.00	0.99	0.64	0.45
Best	Gaussian noise	1.00	0.99	0.95	0.68
Best	Erroneous identification data	1.00	0.99	0.92	0.71
	45 extra spin systems				
Best	None	1.00	1.00	0.96	0.69
Good	None	1.00	0.99	0.68	0.45
Fair	None	1.00	0.92	0.55	0.32
Best	Gaussian noise	1.00	0.99	0.91	0.67

Reduced side-chain information. The spin-system identification data involves experimental procedures such as (H)CCH and (H)C(C$_{\alpha}$CO)NH. These experiments are generally of lower sensitivity than the four triple-resonance data sets used (see Table 1), but are the most sensitive for obtaining side-chain information. Thus, when sensitivity of these experiments deteriorates with increasing rotational correlation time, one expects loss of this type of data first. Thus, in addition to the ''best'' spin-system identification set described above, we also defined ''good'' and ''fair'' identification data sets as shown in Table 2, to simulate such sensitivity losses. In the good set, AMX, AMPX, and long side-chain systems are not distinguishable from each other. The fair set was derived from the good set by randomly changing 20% of the members of the A, G, S, T, and {ILV} classes to the class ''unknown.'' When even fair information cannot be obtained, identities can still be ascertained by selective labeling of several judiciously chosen residues [e.g., (*20*)]. This latter situation was not simulated here.

The results are summarized in Table 3. Individual computation took between 1 and 20 cpu hours on a Silicon Graphics Power Challenge R8000, depending upon the degree of the degradation and the number of extraneous spin systems. When data were omitted, q was calculated based solely on those spin systems still present in the data set. As shown, the mean-field simulated annealing method is highly robust, and the performance declines slowly and monotonically as the data degrades. The values of the energy and q during the run with Gaussian noise, 20% missing data, and 30 extra

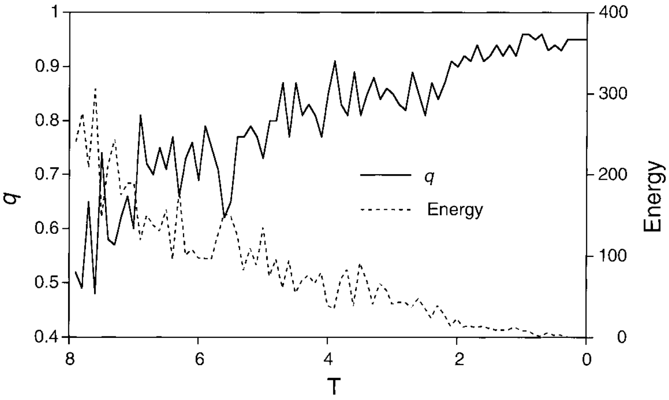

FIG. 1. The values of the energy, calculated with Eq. [2], and q, as defined in Eq. [6], during simulated annealing. The energy values are relative to the energy of the correct assignment. The data from DnaK included the ''best'' side-chain information but was degraded by including Gaussian noise, deleting data for 20% of the spin systems, and adding data for 30 extraneous spin systems from another protein. The run converges to a q value of 0.95, indicating that the assignment was fundamentally correct. The final value of the energy was slightly negative, meaning that given the extensive degradation of the data, the simulated annealing converged to an answer that was actually a better assignment as judged by the cost function than the correct answer would have been.

spin systems are shown in Fig. 1. The strong negative correlation between the energy and q illustrates the validity of the energy function defined in Eq. [2].

Figure 2 shows the values for $\sqrt{P[i, L(i)]}$, where $P[i, L(i)]$ is the occupancy of spin system i at sequence location $L(i)$, for the 197 nonproline residues and extra spin systems at various points during the simulation. (The square root was used to enhance the smaller occupancy values.) The spin systems were sorted with $i = L_c(i)$, so that a correct assignment would correspond to $P[i, L(i)] = \delta_{i,L(i)}$, represented by unit intensity along the $i = L(i)$ diagonal. The extra spin systems are represented past the end of the 167 nonproline residues in the actual DnaK sequence. As shown, the method described handles realistic (thus limited) data from a relatively large protein quite well. Initially, at $T = 8$, the program is quite tentative about the assignments, considering a wide range of possible options. This is highlighted by the band at the top of the plot, representing all of the spin systems that the program considers possibly extraneous. Gradually the program starts to assign spin systems, still avoiding firm commitments. For this reason, the program is able to rectify the small incorrectly assigned regions at $T = 6$ and 4. As the confidence in the assignments grows, the occupancy of the valid DnaK spin systems in the extraneous-spin-system band decreases, and the program can identify and segregate the true extraneous spin systems using the process of elimination, represented by the holonomic constraints. The entries representing missing spin systems become restricted to the missing locations, as all other positions become filled by valid data. As the program cannot distinguish between various omitted spin sets, the various combinations of such spin sets and missing locations form the vast majority of the off-diagonal intensity in the plots. Given the substantial degradation of the data, the final assignment found by the simulated annealing is actually more optimal, given the cost function, than the correct assignment. For instance, extraneous spin system 197 is comfortably assigned to location 151 in the sequence, a position corresponding to omitted data. Likewise, the spin system 121 is more optimally located in sequence location 109, while spin systems 10 and 11 are shifted to identical residue types at sequence locations 11 and 12.

It is of great interest to have some gauge of the reliability of the assignment, measuring what portion is accurate and what portion is not. Performing a simulated annealing run to a final temperature of zero does not provide this information, as it will supposedly yield the best assignment even if only marginally better than other possible assignments. Insight into the distribution of near equally good assignments can be obtained by performing an extended dynamics simulation at a temperature $T = 1$, when the energy scale is set by inversion of the Boltzmann equation. By collecting time-averaged occupations from this run, we can obtain the desired gauge of reliability. The resulting values of $\sqrt{\langle P[i, L(i)]\rangle}$, where $\langle P[i, L(i)]\rangle$ represents the time-averaged value of $P[i, L(i)]$, are shown in Fig. 3. Approximately 86% of the nonomitted data are assigned with an occupancy of greater than 0.95, with 92% assigned with an occupancy of greater than 0.80. All of the incorrect assignments noted above had significant occupancy in other sequence locations, with the sole exception of the assignment of extraneous spin system 197 (occupancy = 0.95). There is of course no computational method that could identify such fortuitous matches without additional information.

DISCUSSION

As mentioned in the Introduction, other researchers have developed automated and semi-automated methods for generating sequence-specific assignments, employing a variety of algorithms including constraint satisfaction, branch-and-bound limited search, genetic, neural net, pseudo-energy minimization, and simulated annealing (*2*). These methods generally use a variety of spectral data, some of which can only be obtained for smaller labeled proteins. Zimmerman and Montelione note that constraint-satisfaction systems, which work by eliminating large portions of combinatorial space of possible solutions have the advantage of only propagating, and not introducing, errors into the assignments (*2*). In contrast, as we demonstrate by the correct convergence of the assignment in cases where identification data incompatible with the proper assignment were included, pseudo-energy-based approaches can actually *rectify* erroneous data by relying on redundancy in the data set. In particular, our

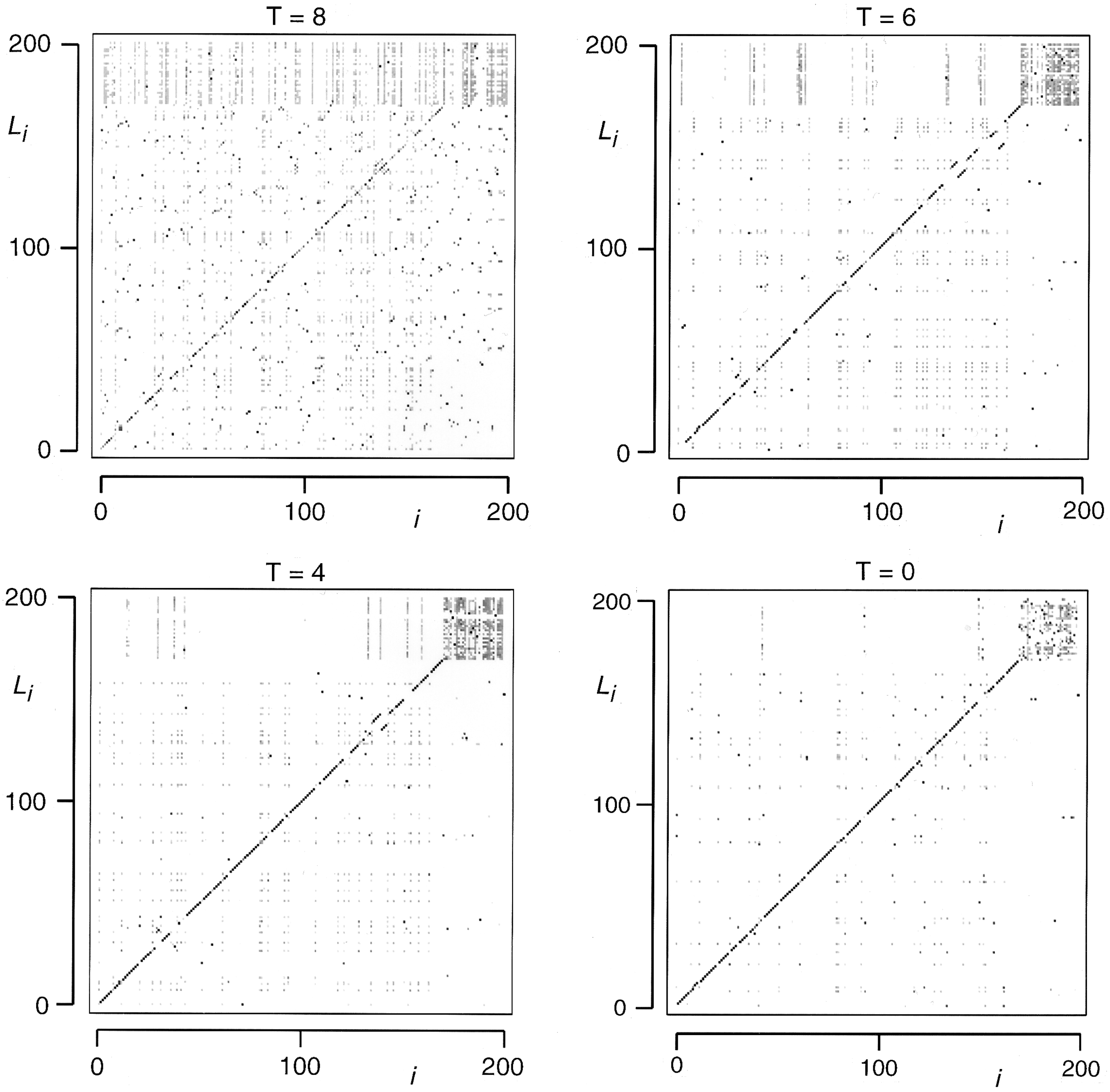

FIG. 2. Density plot of $\sqrt{P[i, L(i)]}$ for the 167 nonproline residues of DnaK and 30 extraneous spin systems from another protein. The square root was used to enhance the smaller occupancy values. The four plots correspond to different stages of the annealing, indicated by the temperature. The data, as described in the legend to Fig. 1, included the ''best'' side-chain information, but was degraded by including Gaussian noise, deleting data for 20% of the spin systems, and adding 30 extraneous spin systems from another protein. The spin systems were sorted so that $i = L_c(i)$, with the additional spin systems appended after the DnaK spin systems. A correct assignment would correspond to intensity only along the $i = L(i)$ diagonal for the first 167 residues. The vast majority of density not along the diagonal corresponds to spin systems whose data were deleted. The speckled region in the upper-right corner of the plot corresponds to the extraneous resonances.

method never eliminates any portion of the combinatorial space of possible solutions because *all* possible assignments are considered probabilistically via a mean-field description. This probabilistic interpretation of assignments leads to a rigorous energetic description via Boltzmann's relation and makes use of a thermodynamic analogy to find the energetic minimum representing the optimal probabilistic assignment. In this way, the pseudo-energy approach can easily include information of various degrees of certainty.

Another concern voiced by Zimmerman is that global optimization generally does not provide partial, yet highly reliable solutions, which are more useful than complete, but uncertain results (*2*). Once again, our mean-field probabilistic approach can provide quantitative measurements of cer-

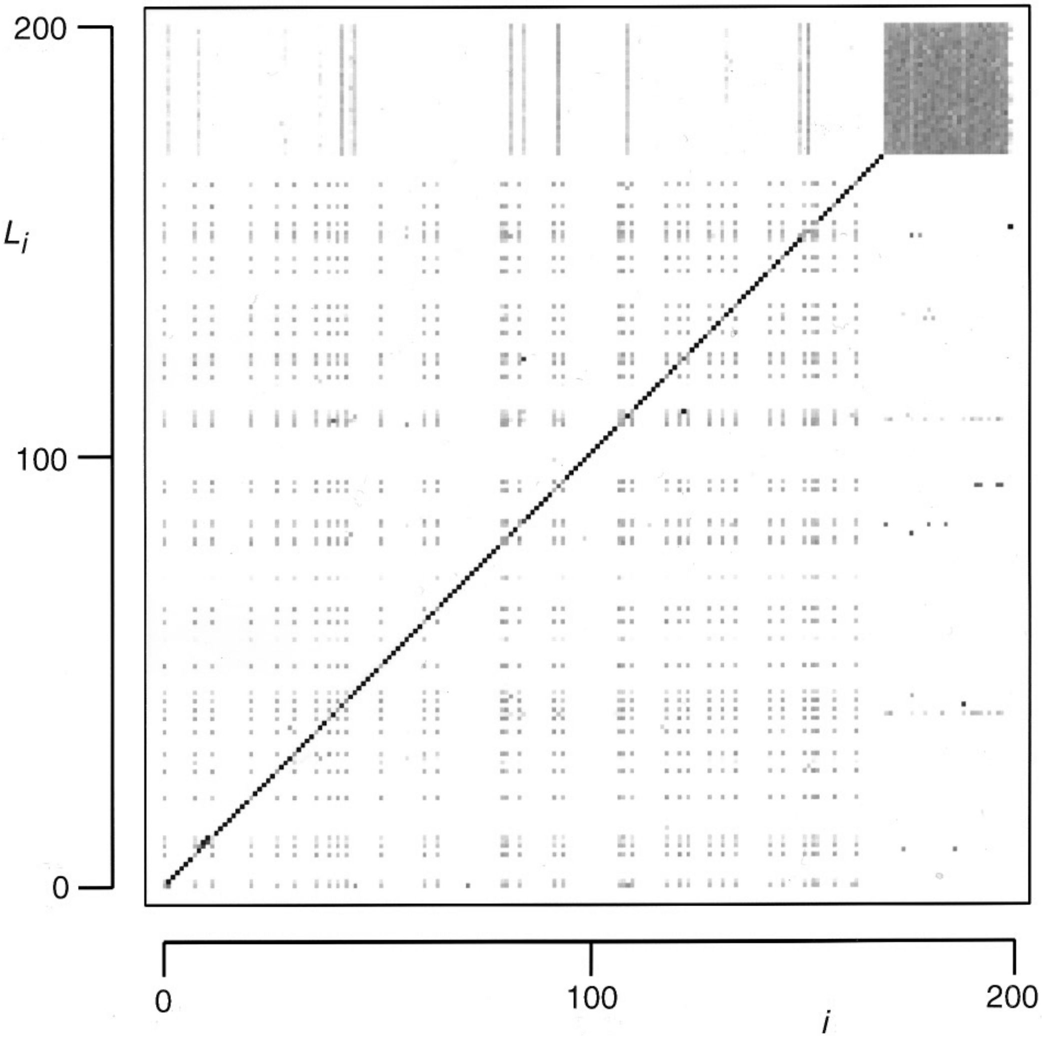

FIG. 3. Density plot of $\sqrt{\langle P[i, L(i)]\rangle}$, where $\langle P[i, L(i)]\rangle$ is the time-averaged value of $P[i, L(i)]$ during a constant temperature run at $T = 1$. Under these conditions, the time average is a quantitative measure of the confidence of the assignment. The data were degraded as described in the legends to Figs. 1 and 2.

tainty, by performing the simulated annealing at a fixed nonzero temperature. Once identified as ambiguous, additional information can be obtained, e.g., from three-dimensional ^{15}N- and ^{13}C-resolved NOE spectra.

We are presenting here an approach rather than a finished software package. As input, our program needs a peak-picked list of 3D NMR resonance data, where one has already established $C_\alpha(i)$, $C_\alpha(i - 1)$, $H_\alpha(i)$, and $H_\alpha(i - 1)$ groupings via amide root resonance connectivities and correlated these spin systems to possible amino acid side-chain identities. The method manifests its strength in its rigor, robustness to uncertainties and errors, simplicity, and generality. The advantage of the use of an energy function is that any source of information, no matter how uncertain, can be included in the computation through the use of the appropriate Boltzmann factor. As such, when more experiments are developed furnishing other sorts of information, these sources can be easily implemented in the theoretical framework. Similarly, more sophisticated techniques for side-chain assignments can be readily included. We have shown how this method works on a protein that is close to the maximum size possible for current solution NMR studies. Further work in developing a more complete software package is in progress.

ACKNOWLEDGMENTS

We thank George Levy and Ravikumar Muppirala for helpful discussions, and Kurt Hillig and James Raines for computational assistance. Financial support was provided by the College of Literature, Science, and the Arts, the Program in Protein Structure and Design, the Horace H. Rackham School of Graduate Studies, NIH Grant GM52421 (to E.Z. and H.W.), and NSF Equipment Grant BIR9512955.

REFERENCES

1. K. Wüthrich, ''NMR of Proteins and Nucleic Acids,'' Wiley, New York, 1986.
2. D. E. Zimmerman and G. T. Montelione, *Curr. Opin. Struct. Bio.* **5,** 664 (1995).
3. R. Bernstein, C. Cieslar, A. Ross, H. Oschkinat, J. Freund, and T. A. Holak, *J. Biomol. NMR* **3,** 245 (1993).
4. R. Wehrens, C. Lucasius, L. Buydens, and G. Kateman, *J. Chem. Inform. Comput. Sci.* **33,** 245 (1993).
5. D. E. Zimmerman, C. Kulikowski, L. L. Wang, B. Lyons, and G. T. Montelione, *J. Biomol. NMR* **4,** 241 (1994).
6. M. S. Friedrichs, L. Müller, and M. Wittekind, *J. Biomol. NMR* **4,** 703 (1994).
7. N. Morelle, B. Brutscher, J.-P. Simorre, and D. Marion, *J. Biomol. NMR* **5,** 154 (1995).
8. J. J. B. Olson and J. L. Markley, *J. Biomol. NMR* **4,** 385 (1994).
9. R. P. Meadows, E. T. Olejniczak, and S. W. Fesik, *J. Biomol. NMR* **3,** 701 (1993).
10. G. Wagner, *J. Biomol. NMR* **3,** 375 (1993).
11. R. A. Venters, C. C. Huang, B. T. Farmer II, R. Troland, L. D. Spicer, and C. A. Fierke, *J. Biomol. NMR* **5,** 339 (1995).
12. L. Verlet, *Phys. Rev.* **159,** 98 (1967).
13. A. Majumdar and E. R. P. Zuiderweg, *J. Magn. Reson. B* **102,** 242 (1993).
14. H. Wang and E. R. P. Zuiderweg, *J. Biomol. NMR* **5,** 207 (1995).
15. H. Wang, ''Application and Development of Multidimensional NMR Spectroscopy in Structural Studies of Oligonucleotides and Proteins,'' Ph.D. dissertation, University of Michigan, 1995.
16. S. Grzesiek and A. Bax, *J. Am. Chem. Soc.* **114,** 6291 (1992).
17. E. T. Olejniczak, R. X. Xu, A. M. Petros, and S. W. Fesik, *J. Magn. Reson.* **100,** 444 (1992).
18. R. T. Clubb, V. Thanabal, and G. Wagner, *J. Biomol. NMR* **2,** 389 (1992).
19. W. Boucher, E. D. Laue, S. Campbell-Burk, and P. J. Domaille, *J. Biomol. NMR* **2,** 631 (1992).
20. S. R. Van Doren, A. V. Kurochkin, Q.-Z. Yei, L. L. Johnson, D. J. Hupe, and E. R. P. Zuiderweg, *Biochemistry* **32,** 13,109 (1993).
21. R. C. Morshauser, H. Wang, G. C. Flynn, and E. R. P. Zuiderweg, *Biochemistry* **34,** 6261 (1995).