J Biomol NMR (2009) 43:97–109
DOI 10.1007/s10858-008-9291-x

ARTICLE

Automated assignment of NMR chemical shifts using peak-particle dynamics simulation with the DYNASSIGN algorithm

Roland Schmucki · Shigeyuki Yokoyama · Peter Güntert

Received: 29 September 2008 / Accepted: 6 November 2008 / Published online: 26 November 2008

Abstract A new algorithm, DYNASSIGN, for the automated assignment of NMR chemical shift resonances was developed in which expected cross peaks in multidimensional NMR spectra are represented by peak-particles and assignment restraints are translated into a potential energy function. Molecular dynamics simulation techniques are used to calculate a trajectory of the system of peak-particles subjected to the potential function in order to find energetically optimal configurations that correspond to correct assignments. Peak-particle dynamics-based simulated annealing was combined with the Hungarian algorithm for local optimization, and a residue-based score was introduced to distinguish between reliable assignments and "unassigned" resonances for which no reliable assignment can be established. The DYNASSIGN algorithm was implemented in the program CYANA and tested with data sets obtained from the experimental NMR data of nine small proteins. With a set of 10 commonly used NMR spectra, on average 82.5% of all backbone and side-chain ^{1}H, ^{13}C and ^{15}N resonances could be assigned with an average error rate of 3.5%.

R. Schmucki · P. Güntert (✉)
Institute of Biophysical Chemistry and Frankfurt Institute for Advanced Studies, Goethe University Frankfurt am Main, Max-von-Laue-Str. 9, 60438 Frankfurt am Main, Germany
e-mail: guentert@em.uni-frankfurt.de

R. Schmucki · S. Yokoyama
Department of Biophysics and Biochemistry, Graduate School of Sciences, The University of Tokyo, Tokyo, Japan

R. Schmucki · S. Yokoyama · P. Güntert
RIKEN Genomic Sciences Center, Yokohama, Japan

P. Güntert
Graduate School of Science, Tokyo Metropolitan University, Hachioji, Tokyo, Japan

Keywords Resonance assignment · Automated assignment · Peak-particle dynamics · DYNASSIGN · CYANA

Introduction

The chemical shift assignment of hydrogen, nitrogen and carbon resonance frequencies is an essential step during the procedure of protein structure determination and studies of protein interactions and dynamics by NMR spectroscopy (Wüthrich 1986). Many of the widely used computer softwares for the calculation of three-dimensional protein structures from NMR data need an as complete as possible set of assigned chemical shifts, in order to extract distance restraints from NOESY spectra via the nuclear Overhauser effect. Nowadays, the analysis of resonance assignments is still often executed manually and requires a considerable amount of time by an experienced spectroscopist. Therefore, the automation of the chemical shift assignment is highly desirable, in particular because other steps of the structure determination procedure, such as peak picking, NOESY cross peak assignment, structure calculation and energy minimization of the resulting structure can already be performed by automated methods, as reviewed recently (Altieri and Byrd 2004; Baran et al. 2004; Gronwald and Kalbitzer 2004; Güntert 2008).

Over the last decade, several methods have been developed to solve the problem of chemical shift assignment in proteins by using computer algorithms, or computer based approaches with manual interaction by a spectroscopist. Most of the automated programs use an analysis scheme which is based on the conventional method (Wüthrich 1986). The principal idea is to, first identify groups of spins that can be correlated by "through-bond"

experiments and establish links to sequential neighbors, and then match segments obtained in this manner onto the primary structure of the protein. Implementations for this approach include, for example, simulated annealing/Monte Carlo algorithms (Hitchens et al. 2003; Leutner et al. 1998), genetic algorithms (Lin et al. 2005), exhaustive search algorithms (Atreya et al. 2000; Coggins and Zhou 2003; Güntert et al. 2000), and heuristic best-first algorithms (Hyberts and Wagner 2003; Zimmerman et al. 1997). Some programs use a combination of algorithms. For instance, the program GARANT (Bartels et al. 1996, 1997) employs a genetic algorithm combined with simulated annealing and local optimization.

In this work, we have investigated a novel approach to solve the chemical shift assignment problem. The principal idea is to interpret the cross peaks expected to occur in NMR spectra as particles moving in a multidimensional simulation space. In our new algorithm, DYNASSIGN, these so-called "peak-particles" are subjected to a potential that is constructed using the information available from the protein sequence and spectra given by the user. In particular, each measured peak in any of the spectra available represents a local minimum of the potential energy function which leads to a mapping of expected peaks onto measured peaks that establishes the assignment. Other terms of the potential function take into account the alignment of peaks containing identical resonances and the chemical shift statistics (Seavey et al. 1991). In analogy to molecular dynamics simulation a peak-particle dynamics algorithm is employed to compute a trajectory of the system of peak-particles according to the laws of classical mechanics in order to find a configuration with minimal energy. During the search of the global energy minimum, peak-particles will drift towards local potential minima represented by measured cross peaks. In order to find configurations with low potential energy faster, the peak-particle dynamics simulation is complemented by a method to reset the position of selected peak-particles periodically in the course of the simulation. Finally, the set of chemical shift assignments with minimal potential energy found constitutes the output of the algorithm.

The DYNASSIGN algorithm was implemented in the program CYANA (Güntert 2003; Güntert et al. 1997) and applied to peak lists obtained from the experimental data of nine proteins with 46 to 90 residues. The results are presented in this paper.

Algorithm

Peak-particles represent expected peaks

In the DYNASSIGN algorithm expected peaks are represented by particles, termed peak-particles, in a D-dimensional space where D is the dimensionality of the spectrum in which a peak is expected to be observed. The expected peaks are generated on the basis of the amino acid sequence of the protein under study and the magnetization transfer pathways of the NMR experiments (Bartels et al. 1996, 1997). Each spectrum constitutes a separate D-dimensional space containing as many particles as peaks are expected to occur. The expected peaks, and hence the peak-particles, are always assigned to the D atoms involved in it. The position coordinates of a peak-particle provide the chemical shift assignment of the resonances involved in the corresponding peak. For instance, in a two-dimensional [^{15}N, ^{1}H]-HSQC spectrum each expected peak involves a proton (^{1}H) and a nitrogen (^{15}N) resonance. Thus, the coordinates of the peak-particle representing a peak in a [^{15}N, ^{1}H]-HSQC spectrum provide the chemical shift assignment of a ^{1}H and a ^{15}N resonance, respectively. In principal, the units of the coordinates are the units for the chemical shifts, namely ppm. However, as described below, each peak-particle coordinate is scaled by a nucleus-specific factor in order to bring the coordinate values of all types of resonances (^{1}H, ^{15}N and ^{13}C) into similar ranges. The formulas below apply to all types of nuclei and spectra that are commonly used with proteins.

The position of a peak-particle n is represented by a $D(n)$-dimensional vector $\boldsymbol{r}_n = (r_n^1,\ldots,r_n^{D(n)})$, where $D(n)$ is the dimensionality of the spectrum in which the peak n, assigned to atoms $\alpha_n^1,\ldots,\alpha_n^{D(n)}$, is expected to occur. To simplify the notation, the collection of all position vectors is described by $\boldsymbol{r} = (\boldsymbol{r}_1,\ldots,\boldsymbol{r}_N)$ with N the total number of peak-particles in the system comprising all available spectra. The measured peaks are numbered from 1 to M, and their locations are described by analogous position vectors.

Potential energy function

The peak-particles are subjected to a potential function $U(\boldsymbol{r})$ that incorporates four essential aspects of the assignment process:

$$U(\boldsymbol{r}) = U_{\text{stat}}(\boldsymbol{r}) + U_{\text{align}}(\boldsymbol{r}) + U_{\text{exist}}(\boldsymbol{r}) + U_{\text{degen}}(\boldsymbol{r}) \qquad (1)$$

We assume that the chemical shift values in proteins follow a Gaussian probability distribution. The mean value and the standard deviation for each ^{1}H, ^{13}C and ^{15}N nucleus in the 20 types of amino acid residues are available from a statistics over a large number of protein chemical shift assignments that are stored in the Biological Magnetic Resonance Data Bank (BMRB) (Seavey et al. 1991). The chemical shift statistics potential $U_{\text{stat}}(\boldsymbol{r})$ accounts for the deviation between the assigned chemical shift and the statistically determined mean value:

$$U_{\text{stat}}(\boldsymbol{r}) = c_{\text{stat}} \sum_{n=1}^{N} p_n \sum_{i=1}^{D(n)} \left(\frac{r_n^i - \omega(\alpha_n^i)}{\sigma(\alpha_n^i)} \right)^2,$$

where $\omega(\alpha_n^i)$ and $\sigma(\alpha_n^i)$ stand for the chemical shift database mean value and standard deviation, respectively, of the atom α_n^i to which the peak-particle n is assigned in dimension i. The summation includes all N peak-particles. Here and in the following, p_n denotes an a priori probability that peak n can be observed in the spectrum. We used $p_n = 1$ for all calculations in this paper. c_{stat} is the overall weight of the chemical shift statistics potential. The effect of this potential term on a peak-particle is illustrated in Fig. 1a.

The peak alignment term $U_{\text{align}}(\boldsymbol{r})$ takes into account that peaks that involve the same atom(s) must be aligned. Thus, the potential is constructed such that there is a force acting on peak-particles which have resonances in common:

$$U_{\text{align}}(\boldsymbol{r}) = c_{\text{align}} \sum_{n=1}^{N} \sum_{i=1}^{D(n)} \sum_{k>n}^{N} \sum_{j=1}^{D(k)} p_n p_k \left(\rho_{nk}^{ij}\right)^2 \delta_{\alpha_n^i \alpha_k^j}$$

The scaled differences $\rho_{nk}^{ij} = (r_n^i - r_k^j)/\Delta_i$ between coordinates of peak-particles n and k are summed over all pairs of peaks that share common atoms, $\alpha_n^i = \alpha_k^j$, using the Kronecker symbol, $\delta_{lm} = 1$ if $l = m$ and $\delta_{lm} = 0$ otherwise. The distances are scaled by the tolerance parameter Δ_i in order to bring contributions from different types of nuclei (^{1}H, ^{13}C, ^{15}N) to similar scales. c_{align} is the overall weight of the peak alignment potential. An illustration of the effect of this potential is given in Fig. 1b.

The peak existence term $U_{\text{exist}}(\boldsymbol{r})$ considers the positions of the measured peaks in the spectra. At the position of each measured peak, a potential well with a negative Gaussian shape is introduced in order to attract the moving peak-particles to the vicinity of a measured peak. Consequently, the scaled distances between peak-particles and measured peak positions are evaluated. Since an expected peak should be close to at least one measured peak, the contributions of all M measured peaks to a single peak-particle n are multiplied:

$$U_{\text{exist}}(\boldsymbol{r}) = c_{\text{exist}} \sum_{n=1}^{N} p_n \prod_{k=1}^{M} {}' q_k \left(1 - e^{-\rho_{nk}/2}\right)$$

The squared distance between an expected peak n and an observed peak k is given by $\rho_{nk} = \sum_{i=1}^{D(n)} \left(\rho_{nk}^{ii}\right)^2$. For each expected peak n, the primed product extends over all peaks observed in the same spectrum as the expected peak n. The weighting factor q_k can be used to adjust the potential well depth for each measured peak k. In this study, these parameters were set to unity. c_{exist} is the overall weight of the peak existence potential. The schematic drawing in

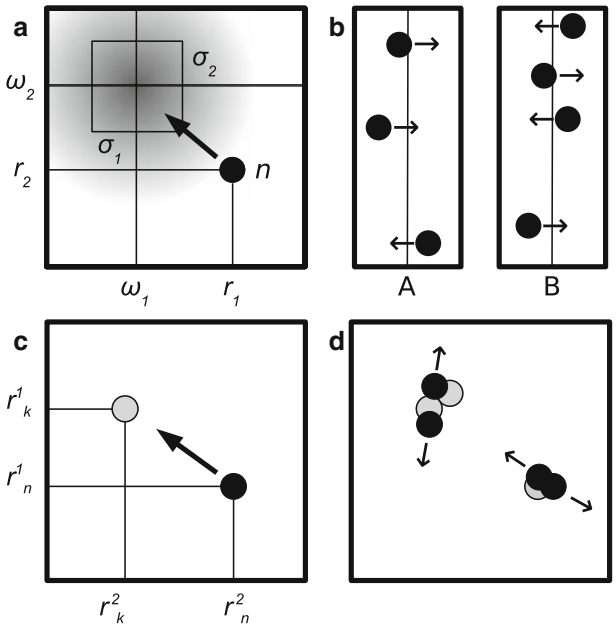


Fig. 1 Schematic representation of the four potential energy terms for peak-particle dynamics simulation, illustrated for the example of a 2D spectrum. Peak-particles are represented by black circles. The positions of measured peaks are indicated by grey circles. **a** Chemical shift statistics potential. A peak-particle n at position (r_1, r_2) experiences a force, indicated by the arrow, in the direction of the statistical mean values ω_1 and ω_2 of the two resonances represented by the peak-particle n. σ_1 and σ_2 are the chemical shift standard deviation for the resonances, respectively, obtained from the chemical shift statistics database. The grey area marks the region where the peak is expected to occur according to the chemical shift statistics. **b** Peak alignment potential. Peak-particles (in the same or different spectra A, B) that share a common atom are subjected to forces (arrows) in order to align their positions in the shared dimension, as indicated by the vertical lines. **c** Peak existence potential. A force pointing to the measured peak k at position (r_k^1, r_k^2) and visualized by the arrow, acts on the peak-particle n at position (r_n^1, r_n^2). **d** Peak degeneracy potential. Small forces depicted by arrows act on peak-particles that approach each other closely. The grey circles indicate measured peaks. The peak degeneracy potential is, however, defined only in terms of the peak-particles and independent from measured peaks

Fig. 1c illustrates the effect of the peak existence potential. Optionally, the range of the existence potential can be defined individually for each measured peak by applying peak-specific chemical shift tolerances Δ. This can be used to assign an attractive potential with longer range for isolated measured peaks than for peaks belonging to a cluster of measured peaks.

The peak degeneracy potential $U_{\text{degen}}(\boldsymbol{r})$ introduces a penalty for degeneracy. Generally, it is unlikely but possible that peaks assigned to different atoms are located at the same position in the spectrum. Consequently, a potential is introduced to induce peak-particles containing different resonances to not occupy the same position. However, in order to allow a certain degree of degeneracy, peak-particles

can adopt positions close to each other. The range of the force is small, and the extent of how close they can approach each other is controlled by the tolerance parameters Δ.

$$U_{\text{degen}}(\boldsymbol{r}) = c_{\text{degen}} \frac{1}{2} \sum_{n=1}^{N} \sum_{k \neq n}^{N} {}' \sum_{i=1}^{D(n)} \left(1 + 2\left|\rho_{nk}^{ii}\right|^3 - 3(\rho_{nk}^{ii})^2\right) \times \theta\left(1 - (\rho_{nk}^{ii})^2\right),$$

where, for each expected peak n, the primed sum runs over all expected peaks k in the same peak list as peak n that do not share any common atoms with peak n. The potential has a polynomial shape. The factor 1/2 stems from the fact that every pair (n, k) of peak-particles is considered twice. c_{degen} is the overall weight of the peak degeneracy potential. Figure 1d illustrates the influence of the peak degeneracy potential U_{degen}.

Potential gradient and force calculation

The interaction of N particles via the potential U can be described by the classical equations of motion. The force $\boldsymbol{f}_i$ acting on a particle i with mass m_i is $\boldsymbol{f}_i = -\nabla_{\boldsymbol{r}_i} U(\boldsymbol{r}_i)$ and the equation of motion can be written as $m_i \ddot{\boldsymbol{r}}_i = \boldsymbol{f}_i$. Since the peak-particles do not have a natural physical mass, we attributed unit masses of 1 kg to all peak particles when simulating the motion of the many-particle system as described in the following. The units of the potential weights c_{stat}, c_{align}, c_{exist}, and c_{degen} are chosen such that the unit of the potential energy is Joule (1 J = 1 kg m^2 s^{-2}). The gradient of the potential of Eq. 1 reads as follows:

$$-\nabla_{\boldsymbol{r}} U(\boldsymbol{r}) = -\nabla_{\boldsymbol{r}} U_{\text{stat}}(\boldsymbol{r}) - \nabla_{\boldsymbol{r}} U_{\text{align}}(\boldsymbol{r}) - \nabla_{\boldsymbol{r}} U_{\text{exist}}(\boldsymbol{r}) - \nabla_{\boldsymbol{r}} U_{\text{degen}}(\boldsymbol{r})$$

In detail, the four terms are

$$\nabla_{\boldsymbol{r}} U_{\text{stat}}(\boldsymbol{r}) = c_{\text{stat}} 2 \sum_{n=1}^{N} p_n \sum_{i=1}^{D(n)} \frac{r_n^i - \omega(\alpha_n^i)}{\sigma(\alpha_n^i)^2} \hat{\boldsymbol{e}}_i$$

$$\nabla_{\boldsymbol{r}} U_{\text{align}}(\boldsymbol{r}) = c_{\text{align}} 2 \sum_{n=1}^{N} \sum_{i=1}^{D(n)} \sum_{k>n}^{N} \sum_{j=1}^{D(k)} p_n p_k \frac{\rho_{nk}^{ij}}{\Delta_i} \delta_{\alpha_n^i \alpha_k^j} \hat{\boldsymbol{e}}_i$$

$$\nabla_{\boldsymbol{r}} U_{\text{exist}}(\boldsymbol{r}) = c_{\text{exist}} \sum_{n=1}^{N} p_n \sum_{l=1}^{M} e^{-\rho_{nl}/2} \sum_{i=1}^{D(n)} \frac{\rho_{nl}^{ii}}{\Delta_i} \hat{\boldsymbol{e}}_i \prod_{k \neq l}^{M} {}' \times q_k \left(1 - e^{-\rho_{nk}/2}\right)$$

$$\nabla_{\boldsymbol{r}} U_{\text{degen}}(\boldsymbol{r}) = c_{\text{degen}} 3 \sum_{n=1}^{N} \sum_{k \neq n}^{N} {}' \sum_{i=1}^{D(n)} \frac{\rho_{nk}^{ii}}{\Delta_i} \left(\left|\rho_{nk}^{ii}\right| - 1\right) \times \theta\left(1 - (\rho_{nk}^{ii})^2\right) \hat{\boldsymbol{e}}_i$$

Here, $\hat{\boldsymbol{e}}_i$ stands for a unit vector pointing in the direction of dimension i.

Integration time step

The Verlet algorithm (Allen and Tildesley 1987), which is widely used in molecular dynamics simulation, is applied for numerically integrating the equations of motion. The method is based on positions $\boldsymbol{r}(t)$, accelerations $\boldsymbol{a}(t) = \ddot{\boldsymbol{r}}(t)$, and the positions $\boldsymbol{r}(t - \Delta t)$ from the previous time step. The equation for advancing the positions is

$$\boldsymbol{r}(t + \Delta t) = 2\boldsymbol{r}(t) - \boldsymbol{r}(t - \Delta t) + \Delta t^2 \boldsymbol{a}(t)$$

The variable t denotes the time and Δt the integration time step. The velocities $\boldsymbol{v}(t) = \dot{\boldsymbol{r}}(t)$needed to evaluate the kinetic energy are obtained from the finite difference formula

$$\boldsymbol{v}(t) = \frac{\boldsymbol{r}(t + \Delta t) - \boldsymbol{r}(t - \Delta t)}{2\Delta t}$$

To perform the integration time step and the kinetic energy calculation, the chemical shift coordinates of the peak-particles are scaled by a nucleus-specific factor in order to bring the scaled coordinate values of all types of resonances (^{1}H, ^{15}N and ^{13}C) into similar ranges. This is achieved by multiplying the chemical shift coordinates of the peak-particles by scaling factors of 1.4 m ppm^{-1} for protons, 0.2 m ppm^{-1} for carbons, and 0.1 m ppm^{-1} for nitrogens. After each simulation step, the inverse scaling yields the new unscaled coordinates of the peak-particles, which are used for the potential and force calculation.

To control the temperature in a simulated annealing schedule, the velocities are scaled in each peak-particle dynamics time step by a factor that depends on the current kinetic temperature and the given reference temperature (Berendsen et al. 1984). This method forces the system towards the desired temperature at a user defined rate, while only slightly perturbing the forces on each particle. The ratio of the integration time step to the coupling constant for weak coupling to a heat bath was set to 0.03.

Consensus chemical shifts

In order to dispose of a single chemical shift assignment for each resonance, the corresponding coordinate of every peak-particle that is assigned to a given resonance is collected to form an ensemble of raw chemical shift assignments. Then, a consensus chemical shift assignment, introduced in the protein structure calculation algorithm FLYA (López-Méndez and Güntert 2006), is determined. The most highly populated chemical shift value in the ensemble is computed for each resonance and selected as the consensus chemical shift value. The consensus chemical shift for a given resonance is the value ω that maximizes the function $f(\omega) = \sum_j \exp\left(((\omega - \omega(j))/\Delta)^2/2\right)$, where the sum runs over all chemical shift values $\omega(j)$ for the given resonance in

the ensemble of raw chemical shift assignments, and Δ denotes the chemical shift tolerance.

Peak-particle dynamics simulation

The operational sequence of the program is presented in the flow chart of Fig. 2. The DYNASSIGN algorithm was implemented in the program CYANA (Güntert 2003; Güntert et al. 1997) using the Fortran 95 programming language. First, input data is read, namely the protein sequence and peaks lists containing all observed peaks. A separate peak list, in XEASY (Bartels et al. 1995) or NMRView (Johnson 2004) format, is read for each spectrum. Then, based on the sequence, a list of all peaks expected to occur in the spectra is generated by a new CYANA command. Next, the peak-particle dynamics simulation algorithm is executed until the exit condition is fulfilled. The output of the DYNASSIGN algorithm is a list containing the chemical shift assignments.

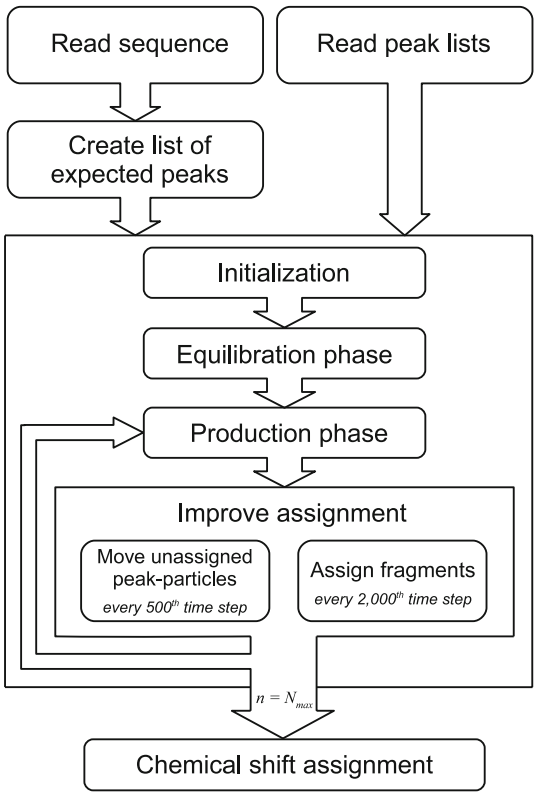


Fig. 2 Flowchart of the DYNASSIGN algorithm. After reading input data and creating a list containing all expected peaks, the peak-particle dynamics simulation is executed over N_{max} steps. Following the initialization, the equilibration and production phases take place. At certain times, the peak-particle dynamics simulation is interrupted to execute a different procedure to improve the assignment. Every 500 steps unassigned peak-particles are moved onto measured peaks, and every 2,000 steps fragments are matched to residues using a complete set of chemical shift assignments. The output of the algorithm comprises a list of chemical shift assignments

Initialization of the peak-particle dynamics simulation

After reading all input data and the user defined parameters, the peak-particles are distributed randomly in the simulation space. For this initial configuration, the force acting on each peak-particle is evaluated. Initial velocities are assigned to each peak-particle by choosing a uniformly distributed random number lying within a nucleus-dependent range of initial velocities for each coordinate.

The algorithm uses cutoffs and the concept of Verlet neighbor lists (Allen and Tildesley 1987) which contain information about the current peak-particle configuration in order to accelerate the evaluation of the potential function and its gradient. Interactions between peak-particles that were further apart than the nucleus-specific cut-off values of 0.3 ppm for protons, 2 ppm for carbons and 4 ppm for nitrogens were neglected when calculating the existence potential. The analogous cut-off values for the degeneracy potential were 0.1 ppm for protons, 0.14 ppm for carbons and 0.08 ppm for nitrogens. The existence and degeneracy potential functions employ distinct pair lists that are initialized at the beginning of the peak-particle dynamics simulation. The distance between the current and previous position of a peak-particle is evaluated and used to decide whether an update is necessary. If there was a large alteration since the last update, e.g. if the distance exceeds a threshold, then the neighbor list is updated. In case that only slight changes compared to the configuration at the previous update occurred, e.g. the distance is lower than the threshold, no update is needed. For instance, a peak-particle with two proton coordinates that has moved by at least 0.07 or 0.12 ppm will trigger an update of the pair list for the existence or degeneracy potential, respectively.

Equilibration and production phase

Following the initialization, the equilibration and production phases are launched which comprise the main part of the peak-particle dynamics simulation algorithm. The simulation space is searched for the global potential energy minimum (or a configuration nearby) using a simulated annealing protocol in which the temperature serves as a parameter to control the kinetic energy of the system.

At the beginning, a certain number of simulation steps are performed at a constant, high temperature until the system has come to equilibrium. At the end of this equilibration period, all memory of the initial configuration should have been lost. Typically, at the end of the equilibration phase, the system has reached a desired end temperature that will be the starting temperature for the next period, the "production phase". During this phase data will be collected. The operational sequence of both phases is identical except that resetting peak-particle positions is

performed only during the production phase. In each time step, the new forces acting on each peak-particle are calculated to obtain the accelerations, and the new positions and velocities are derived according to the Verlet algorithm.

Subsequently, the pair lists for the existence and degeneracy potentials are updated for each peak-particle in case that the condition for an update is fulfilled (Allen and Tildesley 1987). The update condition depends on the actual state of the system and its difference to the previous configuration as described above. The usage of Verlet neighbor lists accelerates the evaluation of potential gradients significantly.

The next step in the simulation protocol, "Improve assignment", has the aim to accelerate the search for consistent and correct assignments by an external procedure, not related to the peak-particle dynamics simulation. The position of selected peak-particles is modified according to the scheme described below. After resetting the positions of certain peak-particles, the forces, velocities, and accelerations as well as the current kinetic and potential energies are recalculated. Every 100 simulation steps, the consensus chemical shift assignment of the current configuration is determined for each resonance. When the exit condition is fulfilled, then the peak-particle dynamics simulation is stopped, otherwise the next time step is executed.

Scoring function for residues

The potential energy value defined in Eq. 1 can be interpreted as a score by virtue of its definition. However, it is not effective enough to judge the quality of a given single chemical shift assignment. A more powerful method is to score all resonances of a residue as a whole. In this way, every chemical shift assignment of a residue obtains the same score.

The residue score is calculated periodically, and the highest score achieved for each residue is stored until the end of the simulation. First, the consensus chemical shift assignment δ_k is determined for each resonance k of the current peak-particle configuration, as described above. Then, the score $S(n)$ of residue n and its assigned (consensus) chemical shifts δ_k is evaluated with the following formula:

$$S(n) = \left(\frac{m_{\text{int}}}{M_{\text{int}}}\right)^2 \left(\frac{m_{\text{seq}}}{M_{\text{seq}}}\right)^2 \prod_k \exp\left(-q\left(\frac{\delta_k - \omega_k}{\sigma_k}\right)^2\right) \times (1 - m_k/M_k)$$

The product runs over all resonances k belonging to residue n. This definition of the score for a residue takes four aspects into account. The first two factors consider how well the chemical shifts correspond to the cross peaks in the spectra. The number of expected intra-residual (sequential) peaks that can be matched to a measured peak is denoted by m_{int} and m_{seq}, respectively, and M_{int} and M_{seq} are the total numbers of intra-residual and sequential peaks. The third term compares the assigned chemical shifts with the chemical shift statistics. δ_k represents the chemical shift value assigned to resonance k, ω_k the mean value and σ_k the standard deviation taken from the chemical shift statistics. The parameter q was set to 0.003. The last term of $S(n)$ takes into account the degeneracy of the chemical shift assignments. Here, m_k is the number of resonances that are degenerate with resonance k and M_k is the total number of resonances of the same type (^{1}H, ^{13}C, or ^{15}N) as resonance k in residue n. Degeneracy occurs when the degeneracy condition $|\delta_l - \delta_k| \leq \Delta/10$ is fulfilled, where δ_l and δ_k denote the chemical shifts assigned to resonances l and k, respectively, and Δ the chemical shift tolerance parameter. The score can take values $0 \leq S(n) \leq 1$, where values near one indicate highly reliable assignments, and unreliable assignments have scores close to zero.

Next, the best (maximal) score $S(n)^{\max}$ and the set of corresponding chemical shift assignments for residue n achieved so far is stored. As the best assignments represented by the maximal score $S(n)^{\max}$ are not necessarily consistent anymore with the best assignments of adjacent residues, a consistency check is performed. Residues whose chemical shifts are inconsistent with the assignments of the two sequentially neighboring residues are penalized by a factor 1/2: $S(n)^{\max} \rightarrow 0.5S(n)^{\max}$.

The final set of chemical shift assignments from the DYNASSIGN algorithm is obtained as follows: The potential energy of the chemical shift configuration with maximal score (which was stored, see above) is evaluated using Eq. 1. If its value is lower than the potential energy of a previously obtained configuration then the current set of chemical shift assignment is stored as final assignment until another set with lower potential energy is found.

Improvement of assignment by resetting peak-particle positions

The potential energy surface is composed of rather sharp local minima at the positions of measured peaks which are separated by comparatively flat areas in-between. Therefore, the system can be trapped in local minima. To partially overcome this problem, selected peak-particles are periodically set to new, energetically favorable positions according to the scheme described below. After re-location of the peak-particles, the peak-particle dynamics simulation continues. There are two different methods to reset the position of the peak-particles.

First, peak-particles which are not in the vicinity of a measured peak are set to the position of a measured peak (see Fig. 3 for a simple illustration). The target measured peak is chosen so that the coordinates of the new position guarantee a consistent (consensus) chemical shift assignment. In this manner, the search for new peak assignments can be accelerated. Additionally, it also releases peak-particles which are trapped in a local energy minimum, or peaks trapped between the potential wells of two or more measured peaks. This method of resetting single peaks is executed every 500 simulation steps.

The second method has the aim to correct backbone assignments locally (within a residue). In contrast to the first method, groups of peaks belonging to the same residue are moved to new, energetically favorable positions. In order to support the peak-particle dynamics simulation finding consistent H_k, N_k, CA_k, CB_k, N_{k+1}, and H_{k+1} shift assignments for a residue k and the following residue $k + 1$, peak-particles are reset to positions which correspond to peaks measured in the spectra for backbone assignment. This local optimization results in correct assignments of H_k, N_k, CA_k, CB_k, N_{k+1}, and H_{k+1} resonances which are consistent within residue k and $k + 1$.

However, the sequential (or global) assignment cannot be achieved solely by this strategy. Thus, the peak-particle dynamics simulation is started again to find a global assignment with low energy. Consequently, peak-particles involving side-chain resonances are automatically pulled towards new locations, consistent with the reset backbone peak-particles, during the following simulation steps. This method of resetting peak-particles is executed every 2,000 steps using the consensus chemical shifts as new coordinates and every 20,000 steps using the chemical shift assignment with maximal residue score as new coordinates.

The groups of peak-particles which are moved simultaneously in the second method are defined as follows. Before the peak-particle dynamics simulation starts, sets of chemical shifts $F_k = \{\delta H_k, \delta N_k, \delta CA_k, \delta CB_k, \delta H_{k+1}, \delta N_{k+1}\}$ (or $F_k^{\mathrm{Gly}} = \{\delta H_k, \delta N_k, \delta CA_k, \delta H_{k+1}, \delta N_{k+1}\}$ for glycine, or $F_k^{\mathrm{Pro}} = \{\delta CA_k, \delta CB_k, \delta H_{k+1}, \delta N_{k+1}\}$ for proline), hereafter named fragments, are extracted from three common backbone spectra: CBCACONH, CBCANH and HN(CA)CO. The obtained fragments are divided into three groups: Fragments containing no CB shifts are identified as fragments belonging to glycine, fragments with missing H and N shifts are identified as fragments belonging to proline, and fragments containing all 6 expected shifts are supposed to belong to the other amino acid types. Next, these fragments have to be matched onto residues in order to improve the chemical shift assignments. In Fig. 4, a simple example is presented. At first, the consensus chemical shift assignment $\delta(i)$ for each resonance i is determined using the raw chemical shift assignments extracted from the current peak-particle configuration, or, every 20,000 steps, from the configuration with maximal residue score.

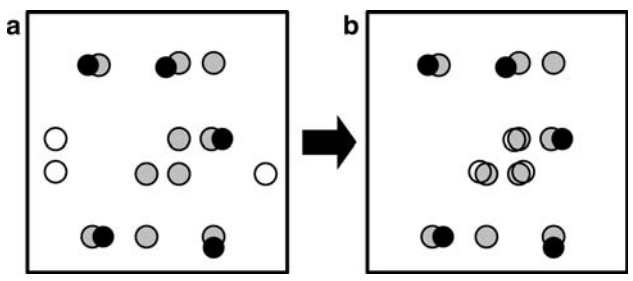
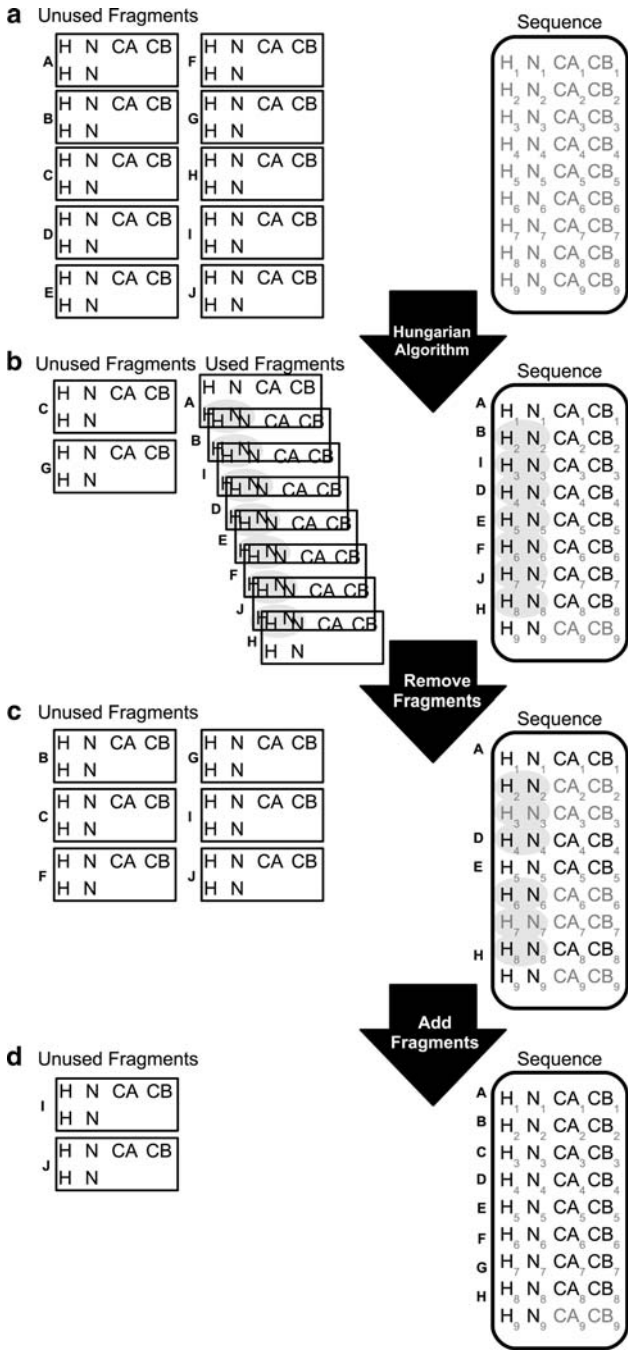


Fig. 3 Illustration of the relocation of unassigned peak-particles. **a** Configuration before resetting peak-particles. **b** New configuration after resetting peak-particles. Peak-particles (black) in the vicinity of measured peaks (grey dots) are not reset. However, peak-particles (circle) which do not have a corresponding measured peak partner are moved onto an unoccupied measured peak

Then, the set of backbone shifts $R_n = \{\delta H_n, \delta N_n, \delta CA_n, \delta CB_n, \delta H_{n+1}, \delta N_{n+1}\}$ of each residue n (Pro and Gly residues are excluded in this stage) is assigned to a fragment F_k using the well-known, polynomial-time "Hungarian algorithm" that is available as subroutine ASSNDX from the CERN Program Library (Bourgeois and Lassalle 1971a, b; Munkres 1957; Silver 1960) (see Fig. 4a, b). This procedure ensures the best fit between the backbone shifts R_n of the current peak-particle configuration (or the configuration with maximal residue score) and the fragments F_k extracted from the backbone assignment spectra. The deviation of the individual chemical shifts as a whole is quantified with the cost function $c = \sum_{R_n,F_k} \sum_j \left(\left(\delta_n^j - \delta_k^j\right)/\Delta_j\right)^2$, where Δ_j denotes the tolerance parameter for resonance j. The second summation runs over all resonances j belonging to the residue. Next, as the Hungarian algorithm cannot take into account of the connectivity between adjacent residues, i.e. matching chemical shift assignments of ^{1}H and ^{15}N, residues and its assigned fragments are flagged in case that the assignment is not consistent with the assignments of the adjacent residues. All assignments of the flagged residues are discarded and put into a pool of "unassigned residues". Accordingly, the unused fragments are returned to a pool of "unused fragments". A search algorithm with variable tolerance parameter seeks for not yet assigned residues and unused fragments in the pool which are consistent with the already assigned residues (Fig. 4b, c). The tolerance used within this search algorithm is slightly increased when no additional fragments that match unassigned residues could be found. At this stage, proline and glycine fragments, F_k^{Pro} and F_k^{Gly}, are also matched onto proline and glycine residues, respectively. Consequently, the number of consistently assigned residues and fragments can be increased. When no more fragments can be matched onto the amino acid sequence (Fig. 4c, d), then peak-particles



◀**Fig. 4** Example illustrating how "fragments" are matched onto the sequence of residues. **a** A pool of 10 unused fragments, A–J, containing H_i, N_i, CA_i, CB_i, H_{i+1} and N_{i+1} resonances on the left, and the H, N, CA and CB resonances of a polypeptide sequence with nine residues on the right. For simplicity, glycine and proline residues are not included in this example. Unassigned residues in the sequence are written with grey font. **b** Situation after application of the Hungarian algorithm on unused fragments. Eight fragments are assigned (used fragments in the middle) and matched onto the sequence written with black font (right side). Two fragments, C and G, are left over in the pool of unused fragments. The last residue of the polypeptide chain cannot be assigned to a fragment. Thus, CA and CB of the last residue are written with grey font. At this stage, H and N shifts of an assigned fragment might be inconsistent with the H and N shifts of the fragment assigned to the preceding residue (marked with grey ovals). **c** Consistency check for matching fragments. Six fragments whose H and N shifts do not match (indicated by grey ovals) are removed from the sequence and returned to the pool of unused fragments (left). They will be used in the next stage, when unused fragments will be added to the sequence. The other four fragments match (black font on the sequence) and will be assigned to the resonances of the sequence. Unassigned resonances are drawn with grey font. **d** Final stage, after adding more fragments to the sequence. All but two fragments are used, and all residues are assigned. The two fragments I and J cannot be used for a consistent and complete assignment of the sequence

are distributed in space according to their newly assigned coordinates, obtained from the consensus chemical shift values, and the peak-particle dynamics simulation is restarted.

Preparation of data sets

Nine small proteins with previously determined nearly complete resonance assignments and a well-defined 3D structure were selected to test the correctness and performance of the DYNASSIGN algorithm. All assignments had been deposited in the BMRB in 2002 or later. The chemical shift assignments of these proteins were obtained from the BMRB website (www.bmrb.wisc.edu) (Seavey et al. 1991) and stored in XEASY/CYANA chemical shift lists. The assignments of the aliphatic ^{1}H and backbone amide ^{1}H resonances were 78–100% complete. Peak lists for CBCANH, CBCACONH, HNCA, HN(CO)CA, HBHACONH, HNHA, HNHB, [^{13}C,^{1}H]-HSQC, CCONH, and HCCH-TOCSY spectra were generated for the given protein sequences on the basis of the experimental chemical shift lists. Peaks involving resonances for which no chemical shift assignment was available from the BMRB were excluded.

Results and discussion

The DYNASSIGN algorithm was tested on nine small proteins with 46–90 residues. The test proteins were selected according to the criteria stated above. Details about these proteins are given in Table 1. For the determination of ^{1}H, ^{13}C and ^{15}N backbone and side-chain chemical shift assignments, a set of 10 common spectra was used: CBCANH, CBCACONH, HNCA, HN(CO)CA, HBHACONH, HNHA, HNHB, [^{13}C,^{1}H]-HSQC, CCONH, and HCCH-TOCSY.

The input data comprised the amino acid sequence and a peak list for each spectrum. The correctness of the chemical shift assignment was assessed by comparison with the

Table 1 List of proteins to which the DYNASSIGN algorithm was applied

Name (Reference)	Acronym[a]	BMRB[b]	PDB	Residues[c]	Pro	Gly
Crambin in DPC micelles (Ahn et al. 2006)	Crambin	6504	1YV8	46	5	4
UBA domain of p62 (Long et al. 2008)	UBA	15591	2JYZ	52	3	5
Chitin-binding domain of *Streptomyces griseus* chitinase (Akagi et al. 2006)	ChiC	10005	2D49	54	1	7
ygdR protein from *E.coli*	ygdR	15079	2JN0	43 (54)[d]	1	4
Second SH3 domain of adaptor Nck (Hake et al. 2008)	Nck	15349	2JS0	54	2	7
Ovomucoid third domain (Song et al. 2003)	Ovomucoid	5472	1M8B	56	2	4
Staphylococcus aureus hypothetical protein SAV1430 (Mercier et al. 2006)	ZR18	5844	1PQX	87	3	2
Hypothetical protein Tm1112 from *Thermotoga maritima*	Tm1112	5357	1LKN	89	5	4
Second PDZ domain of X11alpha (Duquesne et al. 2005)	X11alpha	6113	1Y7 N	81 (90)[d]	3	10

[a] The acronym is used to refer to the protein in the text

[b] The reference chemical shift assignment was retrieved from the BMRB website deposited under the given accession code

[c] For each protein, the total number of residues, and the numbers of proline and glycine amino acids are listed

[d] The number in parenthesis is the total number of residues, including residues for which no reference chemical shifts are available

experimental reference chemical shift list from the BMRB. For each protein, 10^6 peak-particle dynamics simulation steps were performed with a time step size for the Verlet algorithm of $\Delta t = 0.05$ s. The chemical shift tolerance parameters were 0.3 ppm for ^{15}N and ^{13}C, and 0.03 ppm for ^{1}H resonances. The numerical values of the potential weights $c_{\text{stat}} = 0.1$ J, $c_{\text{align}} = 0.3$ J, $c_{\text{exist}} = 1.0$ J, $c_{\text{degen}} = 0.02$ J were chosen empirically. The highest weight was given to the peak existence potential, which is the only potential that relates directly to the experimental data of the protein under study. A low weight was used for the peak degeneracy potential that has the purpose to slightly favor non-degenerate assignments without excluding the occasional degeneracies that occur in proteins.

The system size depends strongly on the size of the protein and the number of NMR spectra used. For example, in the cases of the smallest and largest proteins, Crambin and Tm1112, respectively, there were 1658 and 4031 peak-particles in the system with 10 spectra.

The resulting chemical shift assignments of the proteins after 10^6 steps of peak-particle dynamics simulation are summarized in Table 2. Assignments with residue score $S(n)^{\max} \geq 0.1$ were classified as "correct" if the deviation between the assigned resonance and the reference chemical shift value was smaller than the corresponding chemical shift tolerance, or "wrong" otherwise. Assignments with a residue score $S(n)^{\max} < 0.1$ were classified as "unassigned". Furthermore, if two side-chain resonances, e.g. HB2 and HB3, are assigned to the same chemical shift within the tolerance, then only one of the resonances is considered as "assigned", the other one is set to "unassigned" (see Table 3 for an example that illustrates this aspect).

The percentages in Table 2 were computed relative to the total number of resonances that were assigned in the

Table 2 Chemical shift assignment statistics

Protein	^{1}H			^{13}C			^{15}N		
	Correct	Wrong	Unassigned	Correct	Wrong	Unassigned	Correct	Wrong	Unassigned
Crambin	146 (69)	9 (4)	57 (29)	105 (83)	3 (2)	19 (15)	35 (81)	1 (2)	7 (16)
UBA	203 (79)	10 (4)	43 (17)	141 (90)	6 (4)	10 (6)	48 (98)	0	1 (2)
ChiC	165 (69)	5 (2)	69 (29)	112 (87)	1 (1)	16 (12)	47 (78)	1 (2)	12 (20)
ygdR	173 (75)	9 (4)	48 (21)	119 (84)	3 (2)	19 (13)	43 (96)	1 (2)	1 (2)
Nck	225 (78)	11 (4)	54 (19)	155 (89)	6 (3)	14 (8)	53 (91)	2 (3)	3 (5)
Ovomucoid	202 (77)	7 (3)	55 (21)	132 (88)	4 (3)	14 (9)	55 (93)	0	4 (7)
ZR18	334 (74)	21 (5)	98 (22)	224 (81)	17 (6)	36 (13)	77 (88)	2 (2)	9 (10)
Tm1112	303 (62)	25 (5)	161 (33)	221 (74)	23 (8)	54 (18)	66 (74)	4 (4)	19 (21)
X11alpha	332 (80)	9 (2)	75 (18)	228 (90)	10 (4)	16 (6)	80 (95)	1 (1)	3 (4)

The numbers in parentheses are percentages relative to the total number of resonances in the reference chemical shift list. Assignments with residue score $S(n)^{\max} \geq 0.1$ are classified as "correct" if the deviation between the assigned resonance and the reference chemical shift value is smaller than the corresponding chemical shift tolerance, or "wrong" otherwise. Assignments with a residue score $S(n)^{\max} < 0.1$ are classified as "unassigned"

Table 3 Example illustrating the treatment of degenerate assignments

Resonance	Chemical shift (ppm)			Classification
	Reference	Assigned	Final	
HB2	1.45	1.44	1.44	Correct
HB3	1.92	1.46	–	Unassigned
HG2	1.31	1.32	1.32	Correct
HG3	2.98	2.67	2.67	Wrong

The chemical shift tolerance is $\Delta = 0.03$ ppm. If two resonances are assigned to the same "assigned" shift within the tolerance, then, one of the resonances assignments is "correct" and its "final" shift remains unchanged, whereas the other assignment is classified as "unassigned". If the assigned shifts are not equal within the tolerance, then both shifts are kept but one of it is "wrong"

reference chemical shift list from the BMRB and that gave rise to peaks in the spectra considered. On average 74% of the ^{1}H resonances were assigned correctly. In addition, correct assignments were obtained on average for 85% of the ^{13}C and 88% of the ^{15}N resonances. The number of wrongly assigned resonances is on average 3.5% of the total number of resonances and less than 6% in all cases. The remaining resonances remained unassigned. The residue score $S(n)^{\max}$ thus enables in the majority of cases a clear distinction between reliable assignments and unreliable "unassigned" resonances. The number of unassigned ^{1}H resonances varies between 17% and 33% among the nine test proteins. The fraction of unassigned ^{13}C and ^{15}N resonances is considerably lower (2–21%) than the fraction of unassigned ^{1}H resonances. Additionally, the error ratio is better for ^{13}C and ^{15}N because many of the wrongly assigned resonances are filtered by the residue score criteria.

In practice, an algorithm, such as the present one, that provides a smaller number of assignments with high reliability is in general preferable over an algorithm that returns an assignment for all, or almost all, resonances without quality assessment because in the latter case the conscientious user is still obliged to manually check and validate every assignment. The completeness of the assignments obtained by DYNASSIGN can be compared with the protein chemical shift assignments deposited in the BMRB (Seavey et al. 1991). Almost all of these assignments have been established manually or by interactive semiautomatic methods. The average completeness of the ^{1}H assignments is 88% for the 2,645 entries for proteins larger than 4 kDa, excluding obviously incomplete entries with less than 50% assignments. The completeness of the chemical shift assignments by the DYNASSIGN algorithm is thus only slightly lower than that of typical manual assignments of proteins. On the other hand, we have shown earlier that 3D protein structures can be obtained with an accuracy of about 1 Å backbone RMSD by the fully automated FLYA approach on the basis of chemical shift assignments that are 85% identical to those obtained manually (López-Méndez and Güntert 2006), and that a similar extent of chemical shift assignments enables robust protocols for combined automated NOESY assignment and structure calculation with CYANA (Jee and Güntert 2003). Obviously, the results of the automated DYNASSIGN algorithm can be transferred to one of the software packages for interactive NMR spectrum analysis in order to extend the assignments before proceeding to the collection of conformational restraints and the structure calculation.

In Fig. 5, the energy and the ratio of correct resonance assignments as a function of simulation time are presented. The data originates from the protein ChiC (Akagi et al. 2006). The first 10,000 time steps belong to the equilibration phase at a constant high temperature, which is followed by the production phase during which the temperature is gradually decreased up to a total number of $N = 10^6$ simulation steps. Furthermore, the position of the peak-particles is reset periodically which results in numerous spikes but also a more rapid overall decrease of the total energy (Fig. 5a).

The plot in Fig. 5b displays the individual terms of the potential energy. It can be seen that the peak existence potential (red) and the chemical shift statistics potential (blue) give similar contributions to the total energy that are larger than the contributions from the peak alignment potential (green) and the peak degeneracy potential (magenta). This is due to the choice of the potential weights, $c_{\text{stat}} = 0.1$ J, $c_{\text{align}} = 0.3$ J, $c_{\text{exist}} = 1.0$ J, and $c_{\text{degen}} = 0.02$ J. Furthermore, the alignment of peak-particles is achieved of necessity through resetting peak-particle positions. The peak degeneracy potential has the lowest weight, $c_{\text{degen}} = 0.02$, and applies only to a small number of peaks compared to the other potential terms that apply to virtually all peaks. Thus, the degeneracy potential remains small throughout the equilibration and production phase. In contrast, the existence potential has highest weight, $c_{\text{exist}} = 1.0$, and operates on all peaks. Therefore, it shows the biggest change during the simulation. The weighting factor $c_{\text{stat}} = 0.1$ J for the chemical shift statistics potential is set to a small value in order to allow also assignments that deviate strongly from the mean values of the chemical shift statistics. But it is effective, together with the peak existence potential, to hinder the peak-particles from moving too far away from the area where peaks are expected to occur. Note that there are no boundary conditions included in the peak-particle dynamics simulation.

In Fig. 5c the percentage of correctly assigned resonances with maximal residue score $S(n)^{\max}$ is displayed. The ^{15}N resonances (green) show the best result, followed by the ^{13}C resonances (blue). The fraction of correctly

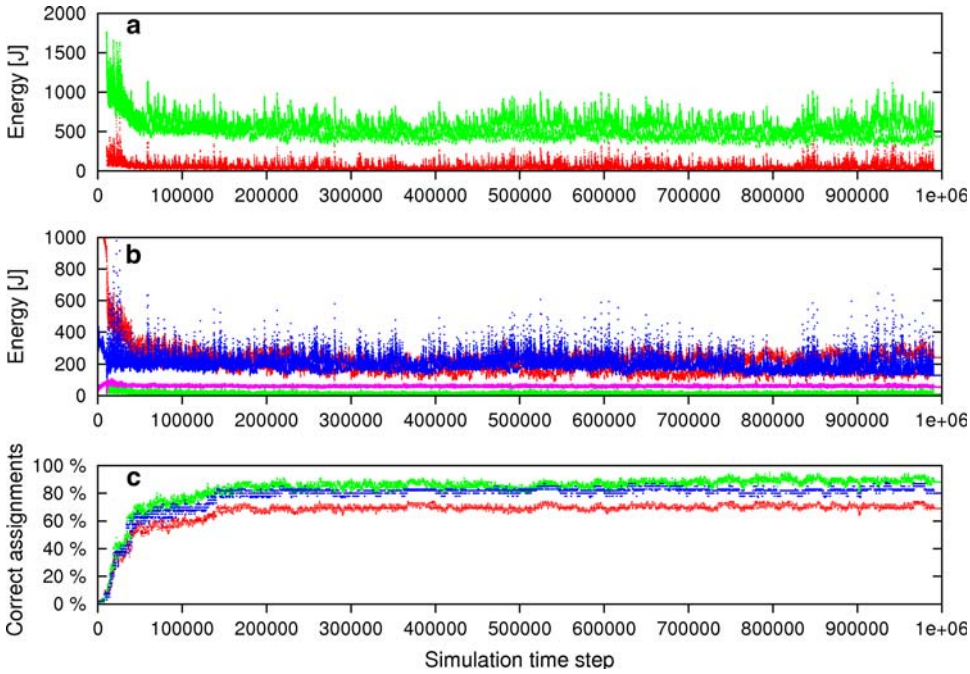


Fig. 5 Plots of energy terms and of the ratios of correct assignment during the 10^6 steps of peak-particle dynamics simulation for the protein ChiC (Akagi et al. 2006). **a** Total energy (green) and kinetic energy (red). Most energy values for the equilibration phase are off-scale. **b** Chemical shift statistics potential (blue), peak alignment potential (green), peak existence potential (red) and peak degeneracy potential (magenta). **c** Percentage of correctly assigned ^{1}H (red), ^{13}C (blue) and ^{15}N (green) resonances with maximal residual score $S(n)^{\max}$ is shown

assigned ^{1}H resonances (red) is lower than the ones for ^{15}N or ^{13}C because the many ^{1}H side-chain resonances are more difficult to assign than backbone resonances. After 10% of the total simulation time approximately half of the resonances are assigned correctly, and only slight improvements of the chemical shift assignments are seen after 20% of the simulation time of this deliberately long run. Note that the progress of correct chemical shift assignments is in agreement with the development of the potential energy, shown in Fig. 5a, b.

The DYNASSIGN algorithm performs a peak-particle dynamics simulation of freely moving peak-particles representing the cross peaks in the NMR spectra. It is interesting to visualize the individual peak-particles during the simulation. For this purpose, the coordinates of all peak-particles in the [^{13}C,^{1}H]-HSQC spectrum were recorded at three different times of the same peak-particle dynamics run as in Fig. 5 and graphically presented as snapshots in Fig. 6. All peaks are considered and displayed, including those originating from resonances that were classified as "unassigned" in Table 2. The reference peak position is marked with a square (□) whereas for the peak-particles itself a filled circle (○) is used when its position is correct and a cross sign (×) when its position is wrong. Figure 6a shows a snapshot of the configuration at simulation step $n = 5{,}000$ during the initialization phase. It is not surprising that the majority of the peak-particles are not on their correct position because the peak-particles started from randomly distributed positions and move with relatively high velocity during this phase. Next, the equilibration phase is launched. Figure 6b shows the configuration at simulation step $n = 100{,}000$. Most of the peak-particles have found a corresponding measured peak (reference peak) although not necessarily the correct one yet. As the simulation is evolving, only slight changes can be observed. The final configuration after 10^6 simulation steps is presented in Fig. 6c.

All calculations were performed on a single Intel processor with 2.4 GHz clock frequency running under a Linux operating system. The typical computation time was a few hours for a small protein, depending mainly on the number of peak-particles. In other words, the size of the protein and the number and type of NMR spectra used are determining the computation time.

Conclusion and outlook

In this paper we have developed a new approach for solving the NMR chemical shift assignment problem that is different from all earlier methods that have been developed for this purpose. Test calculations with a series of small proteins have shown that our algorithm is capable to automatically assign backbone and side-chain chemical shifts. In the test calculations correct assignments could be distinguished from wrong ones using a residue-wise scoring function such that the extent of wrong assignments is on average below 4% for the data sets used in this study. This provides a proof of principle for the new method. Our method is general in that peak lists from any set of spectra can be used for which the magnetization transfer pathways for generating the expected peaks have been defined in a library.

On the other hand, there are limitations that will have to be overcome by further research. In more difficult cases, a

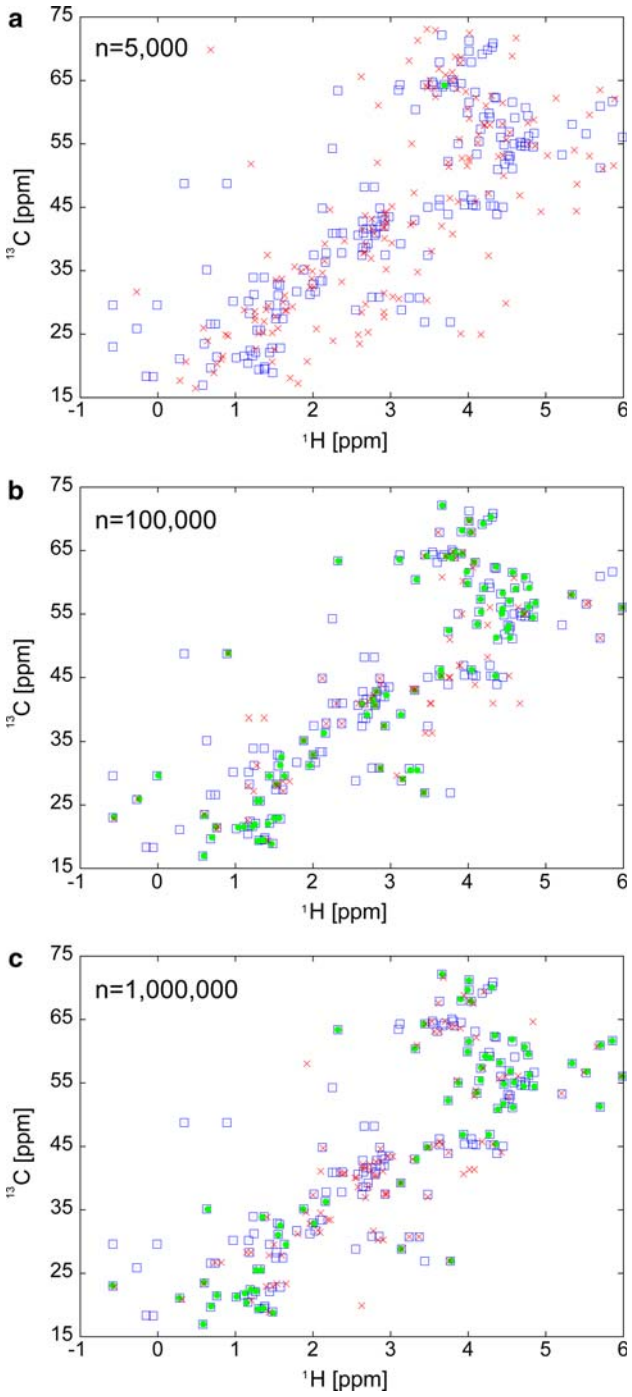


Fig. 6 Snapshots of the peak-particle positions and the observed peaks in the [^{13}C,^{1}H]-HSQC spectrum of the protein ChiC (Akagi et al. 2006) at three different times. **a** After 5,000 steps of peak-particle dynamics simulation. **b** After 100,000 steps. **c** After 1,000,000 steps. Observed peaks are marked with a square (□), correctly positioned peak-particles with a filled circle (○), and wrongly positioned peak-particles with a cross sign (×). In this representation, all peaks are displayed, including peaks involving resonances with residual score $S(n)^{\max} < 0.1$, which are classified as "unassigned" in Table 2

significant part of the resonances can remain unassigned. An improvement of the efficiency of the algorithm in terms of the extent of assignments, the robustness against imperfections of the peak lists, and the computation time is needed for realistic applications. The computation time increases with the number of peak-particles, i.e. the size of the protein and the number of spectra. A more efficient implementation of the potential and gradient computation is conceivable by algorithmic improvements and parallelization. While the original idea of treating the resonance assignment problem by peak-particle dynamics-driven simulated annealing is attractive from a basic point of view, it is in practice important to combine peak-particle dynamics simulation with resetting peak-particle positions by the heuristic algorithm described in the paper. Test runs without "matching fragments onto residues" and relocating individual peak-particles yielded a significantly lower number of correctly assigned chemical shifts. Similar observations had been made with the GARANT program that combines a genetic algorithm with a heuristic local optimization method (Bartels et al. 1996, 1997). As the method of "matching fragments onto residues" applies in our algorithm only to selected backbone atoms but does not include side-chain resonances, one could consider combining the peak-particle dynamics simulation approach with an external backbone assignment algorithm. While the backbone resonances would be assigned mainly by means of the other algorithm, side-chain assignments could be determined by the peak-particle dynamics simulation method.

Acknowledgements Financial support by the National Project on Protein Structural and Functional Analyses of the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT), the Lichtenberg program of the Volkswagen Foundation, and a Grant-in-Aid for Scientific Research of the Japan Society for the Promotion of Science (JSPS) is gratefully acknowledged.

References

Ahn HC, Juranic N, Macura S, Markley JL (2006) Three-dimensional structure of the water-insoluble protein crambin in dodecylphosphocholine micelles and its minimal solvent-exposed surface. J Am Chem Soc 128:4398–4404

Akagi K, Watanabe J, Hara M, Kezuka Y, Chikaishi E, Yamaguchi T, Akutsu H, Nonaka T, Watanabe T, Ikegami T (2006) Identification of the substrate interaction region of the chitin-binding domain of Streptomyces griseus chitinase C. J Biochem 139:483–493

Allen MP, Tildesley DJ (1987) Computer simulation of liquids. Clarendon, Oxford

Altieri AS, Byrd RA (2004) Automation of NMR structure determination of proteins. Curr Opin Struct Biol 14:547–553

Atreya HS, Sahu SC, Chary KVR, Govil G (2000) A tracked approach for automated NMR assignments in proteins (TATAPRO). J Biomol NMR 17:125–136

Baran MC, Huang YJ, Moseley HNB, Montelione GT (2004) Automated analysis of protein NMR assignments and structures. Chem Rev 104:3541–3555

Bartels C, Xia TH, Billeter M, Güntert P, Wüthrich K (1995) The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. J Biomol NMR 6:1–10

Bartels C, Billeter M, Güntert P, Wüthrich K (1996) Automated sequence-specific NMR assignment of homologous proteins using the program GARANT. J Biomol NMR 7:207–213

Bartels C, Güntert P, Billeter M, Wüthrich K (1997) GARANT—A general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. J Comput Chem 18: 139–149

Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. J Chem Phys 81:3684–3690

Bourgeois F, Lassalle JC (1971a) Algorithm 415: algorithm for the assignment problem (rectangular matrices). Commun ACM 14:805–806

Bourgeois F, Lassalle JC (1971b) An extension of the Munkres algorithm for the assignment problem to rectangular matrices. Commun ACM 14:802–804

Coggins BE, Zhou P (2003) PACES: protein sequential assignment by computer-assisted exhaustive search. J Biomol NMR 26:93–111

Duquesne AE, Ruijter M, Brouwer J, Drijfhout JW, Nabuurs SB, Spronk CA, Vuister GW, Ubbink M, Canters GW (2005) Solution structure of the second PDZ domain of the neuronal adaptor X11alpha and its interaction with the C-terminal peptide of the human copper chaperone for superoxide dismutase. J Biomol NMR 32:209–218

Gronwald W, Kalbitzer HR (2004) Automated structure determination of proteins by NMR spectroscopy. Prog Nucl Magn Reson Spectrosc 44:33–96

Güntert P (2003) Automated NMR protein structure calculation. Prog Nucl Magn Reson Spectrosc 43:105–125

Güntert P (2008) Automated structure determination from NMR spectra. Eur Biophys J. doi:10.1007/s00249-008-0367-z

Güntert P, Mumenthaler C, Wüthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol 273:283–298

Güntert P, Salzmann M, Braun D, Wüthrich K (2000) Sequence-specific NMR assignment of proteins by global fragment mapping with the program MAPPER. J Biomol NMR 18: 129–137

Hake MJ, Choowongkomon K, Kostenko O, Carlin CR, Sonnichsen FD (2008) Specificity determinants of a novel Nck interaction with the juxtamembrane domain of the epidermal growth factor receptor. Biochemistry 47:3096–3108

Hitchens TK, Lukin JA, Zhan YP, McCallum SA, Rule GS (2003) MONTE: an automated Monte Carlo based approach to nuclear magnetic resonance assignment of proteins. J Biomol NMR 25:1–9

Hyberts SG, Wagner G (2003) IBIS—A tool for automated sequential assignment of protein spectra from triple resonance experiments. J Biomol NMR 26:335–344

Jee J, Güntert P (2003) Influence of the completeness of chemical shift assignments on NMR structures obtained with automated NOE assignment. J Struct Funct Genom 4:179–189

Johnson BA (2004) Using NMR view to visualize and analyze the NMR spectra of macromolecules. Meth Mol Biol 278:313–352

Leutner M, Gschwind RM, Liermann J, Schwarz C, Gemmecker G, Kessler H (1998) Automated backbone assignment of labeled proteins using the threshold accepting algorithm. J Biomol NMR 11:31–43

Lin HN, Wu KP, Chang JM, Sung TY, Hsu WL (2005) GANA—a genetic algorithm for NMR backbone resonance assignment. Nucleic Acids Res 33:4593–4601

Long J, Gallagher TR, Cavey JR, Sheppard PW, Ralston SH, Layfield R, Searle MS (2008) Ubiquitin recognition by the ubiquitin-associated domain of p62 involves a novel conformational switch. J Biol Chem 283:5427–5440

López-Méndez B, Güntert P (2006) Automated protein structure determination from NMR spectra. J Am Chem Soc 128:13112–13122

Mercier KA, Baran M, Ramanathan V, Revesz P, Xiao R, Montelione GT, Powers R (2006) FAST-NMR: functional annotation screening technology using NMR spectroscopy. J Am Chem Soc 128:15292–15299

Munkres J (1957) Algorithms for the assignment and transportation problems. J Appl Math 5:23–38

Seavey BR, Farr EA, Westler WM, Markley JL (1991) A relational database for sequence-specific protein NMR data. J Biomol NMR 1:217–236

Silver R (1960) An algorithm for the assignment problem. Commun ACM 3:605–606

Song J, Laskowski M Jr, Qasim MA, Markley JL (2003) Two conformational states of turkey ovomucoid third domain at low pH: three-dimensional structures, internal dynamics, and interconversion kinetics and thermodynamics. Biochemistry 42:6380–6391

Wüthrich K (1986) NMR of proteins and nucleic acids. Wiley, New York

Zimmerman DE, Kulikowski CA, Huang YP, Feng WQ, Tashiro M, Shimotakahara S, Chien CY, Powers R, Montelione GT (1997) Automated analysis of protein NMR assignments using methods from artificial intelligence. J Mol Biol 269:592–610