論文の内容の要旨

Peak-particle dynamics for automated NMR resonance assignment

（ピーク粒子動力学計算を用いたNMR自動帰属法の開発）

シュムキ　ローランド

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 nuclear magnetic resonance spectroscopy (NMR). Many of the widely used computer softwares for the calculation of three-dimensional (3D) 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 process 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.

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 where 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, genetic algorithms, exhaustive search and heuristic best-first algorithms.

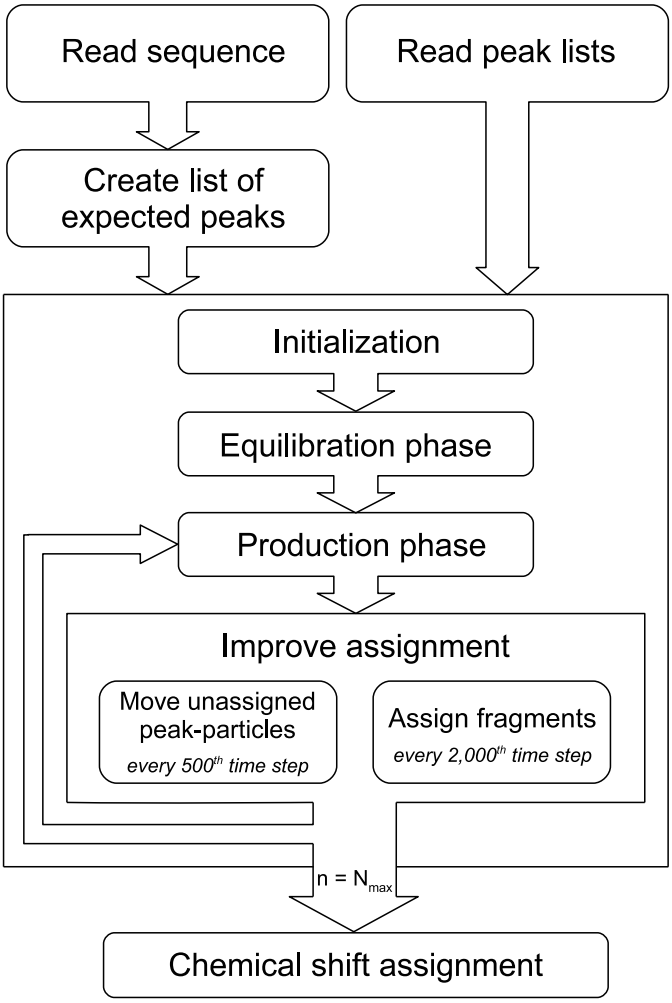

Figure 1 *Flowchart of the DYNASSIGN algorithm. After reading input data and creating a list containing all expected peaks, the peak-particle dynamics simulation is executed over N_{max} steps. Following the initialization, the equilibration and production phases take place. At certain times, the peak-particle dynamics simulation is interrupted and a heuristic algorithm is executed to improve the assignment. Every 500 steps unassigned peak-particles are moved onto measured peaks, and every 2000 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.*

In the first part of the thesis, a novel approach to solve the chemical shift assignment problem has been investigated. The principal idea is to interpret the cross peaks expected to occur in NMR spectra as particles moving in a multidimensional simulation space. In the 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. 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 for 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 heuristic algorithm to reset the position of selected peak-particles periodically in the course of the simulation (Figure 1). Finally, the set of chemical shift assignments with minimal potential energy found constitutes the output of the algorithm.

The DYNASSIGN algorithm was incorporated into the NMR protein structure calculation program CYANA. It was applied to peak lists obtained from the experimental data of nine small proteins with previously determined nearly complete resonance assignments and a well-defined 3D structure. 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 Biological Magnetic Resonance Data Bank (BMRB) were excluded. Test calculations have shown that this algorithm is capable to automatically assign backbone and side chain chemical shifts; 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%. Correct assignments could be distinguished from wrong ones using a residue-wise scoring function. These results provide a proof of principle for the new method. The 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. 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. Test runs without the heuristic algorithm yielded a significantly lower number of correctly assigned chemical shifts. As the heuristic algorithm applies 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.

In the second part of the thesis, a probability function for estimating expected peak overlap in protein NMR spectra is derived with a minimal set of assumptions. Given the primary structure of the

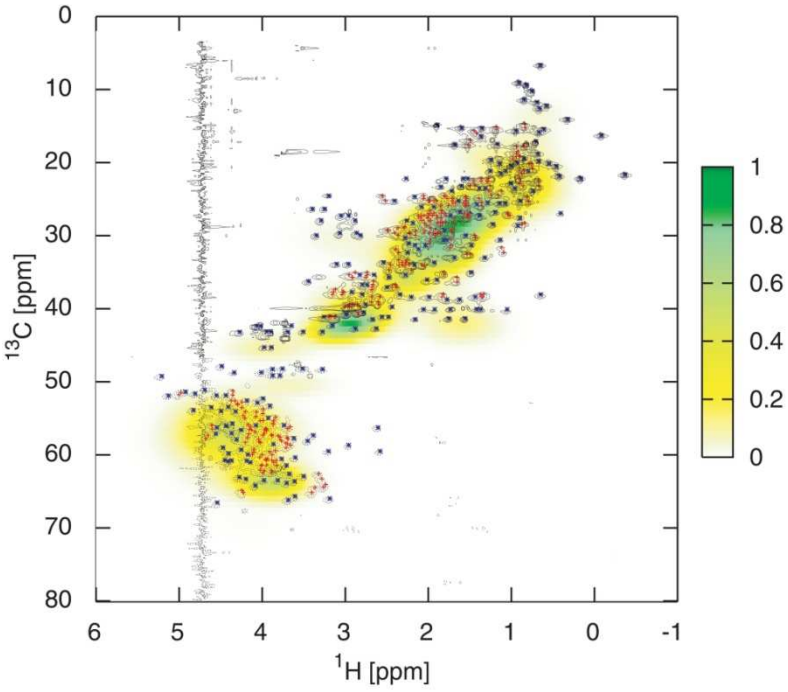

Figure 2 *Plot of calculated peak over-lap for the [^{13}C,^{1}H]-HSQC spectrum superimposed with a contour plot of the real spectrum. Areas with high, medium and low overlap probability are indicated by green, yellow and white color, respectively. Red and blue dots represent real peaks with and without overlap, respectively.*

protein, all theoretically expected cross peaks in an NMR spectrum are determined on the basis of the knowledge of the protein sequence and the magnetization transfer pathways of the pulse sequence that was used to acquire the spectrum. A mathematical model to describe peak overlap is formulated and a parameter which measures the expected number of peaks at a specific position in the spectra is introduced. It is assumed that the atom resonance frequencies (i. e. the peak maxima) are distributed according to a normal Gauss distribution with mean value and standard deviation taken from the chemical shift statistics. The derived formula is an approximation for the theoretically expected peak overlap in NMR spectra. Real data sets from two proteins were used to verify the expected peak overlap probability function. The results are presented in two-dimensional data plots (Figure 2). Peak overlap prediction provides useful information about the expected content of the spectra before performing the NMR experiment. Other possible applications are in computer algorithms for automated resonance frequency assignment.

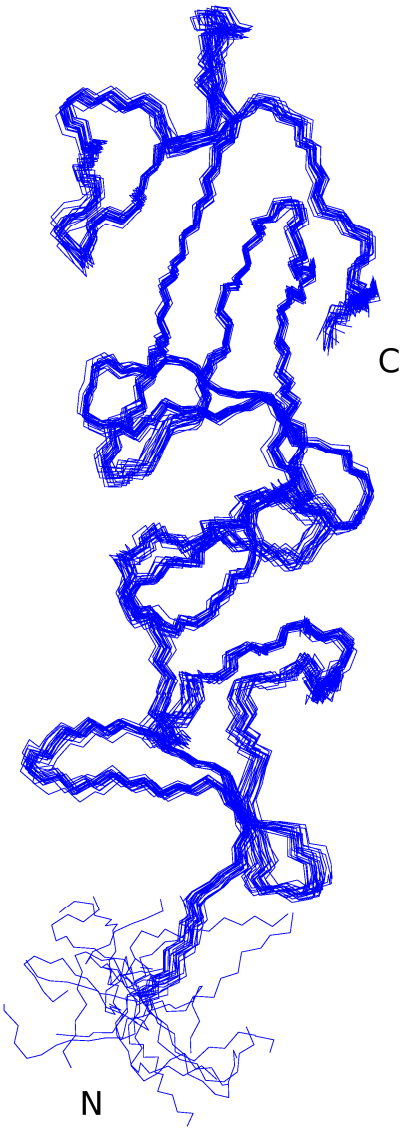

Figure 3 *Line plot of the determined 3D NMR structure of the choline-binding protein. A trace of the backbone atoms for the ensemble of the 20 structures with the lowest CYANA target function is presented.*

In the last part, the 3D NMR solution structure of the *C*-terminal choline-binding domain of the major pneumococcal autolysin harboring a Val317 to Thr mutation was determined and characterized. Autolysin from *Streptococcus pneumoniae* is a bacterial cell wall hydrolase and responsible for cellular autolysis that causes severe infections. The 3D structure was determined by using standard NMR methods with a precision characterized by RMSD values to the mean coordinates of 0.41 Å for the backbone and 0.60 Å for all heavy atoms in the structured region excluding the first seven residues (Figure 3). The obtained tertiary structure is composed of six β-hairpins and choline-binding domains whose primary structures are characterized by the choline-binding repeat motif. Based on the superhelical arrangement of the repeating structural units, the *N*-terminal part can be categorized as a member of the structural family of solenoids. Choline titration experiments were performed for studying the six binding sites; each one is composed of three aromatic and one acidic amino acid residues. According to observations from the titration experiments, at low concentration, the choline binds first in the rigid *N*-terminus, while the overall structure adopts an intermediate state waiting to bind more choline at the remaining sites. Relaxation time measurements revealed that the *N*-terminal part of the protein is rigid and the *C*-terminal is flexible. Moreover, sequence alignment and comparisons with structural homologous proteins are presented and discussed.