J Biomol NMR (2008) 41:127–138
DOI 10.1007/s10858-008-9243-5

ARTICLE

Automated sequence-specific protein NMR assignment using the memetic algorithm MATCH

Jochen Volk · Torsten Herrmann · Kurt Wüthrich

Received: 14 December 2007 / Accepted: 17 March 2008 / Published online: 30 May 2008

Abstract MATCH (*M*emetic *A*lgori*t*hm and *C*ombinatorial Optimization *H*euristics) is a new memetic algorithm for automated sequence-specific polypeptide backbone NMR assignment of proteins. MATCH employs local optimization for tracing partial sequence-specific assignments within a global, population-based search environment, where the simultaneous application of local and global optimization heuristics guarantees high efficiency and robustness. MATCH thus makes combined use of the two predominant concepts in use for automated NMR assignment of proteins. Dynamic transition and inherent mutation are new techniques that enable automatic adaptation to variable quality of the experimental input data. The concept of dynamic transition is incorporated in all major building blocks of the algorithm, where it enables switching between local and global optimization heuristics at any time during the assignment process. Inherent mutation restricts the intrinsically required randomness of the evolutionary algorithm to those regions of the conformation space that are compatible with the experimental input data. Using intact and artificially deteriorated APSY-NMR input data of proteins, MATCH performed sequence-specific resonance assignment with high efficiency and robustness.

Keywords Protein NMR · Sequence-specific resonance assignment · Genetic algorithm · Automation

Introduction

Sequence-specific NMR assignment of polypeptide chains is aimed at obtaining resonance assignments of known chemical structures consisting of a random linear array of multiple copies of the 20 proteinogenic amino acid residues. NMR experiments in common use for resonance assignments identify limited fragments of the polypeptide via scalar couplings. In homonuclear ^{1}H NMR, these fragments represent the intra-residual ^{1}H "spin systems", and sequential assignment of two or several sequentially neighboring spin systems can be achieved using ^{1}H–^{1}H dipolar couplings manifested in nuclear Overhauser effects (NOE) (Wüthrich 1986). In ^{1}H,^{13}C,^{15}N-heteronuclear triple resonance NMR (Montelione and Wagner 1989, 1990; Ikura et al. 1990; Kay et al. 1990), the connected fragments can extend over multiple sequentially adjoining residues, and identification of neighboring fragments is achieved by chemical shift matching of overlapping atoms. Sequence-specific information is obtained from assignment of individual ^{1}H spin systems or heteronuclear fragments to amino acid types, based either on recognizing characteristic peak patterns or on statistical assessments of the chemical shift values. With the thus identified sequence features in the NMR-connected polypeptide segments, these can be matched with the chemically determined amino acid sequence to obtain the sequence-specific assignment (Wüthrich 1983). In principle, an automated procedure seems to be the approach of choice

J. Volk · T. Herrmann · K. Wüthrich (✉)
Institut für Molekularbiologie und Biophysik, ETH Zürich,
CH-8093 Zurich, Switzerland
e-mail: wuthrich@mol.biol.ethz.ch

Present Address:
T. Herrmann
Université de Lyon, CNRS/ENS Lyon/UCB-Lyon 1,
Centre Européen de RMN à Très Hauts Champs de Lyon,
5 rue de la Doua, 69100 Villeurbanne, France

K. Wüthrich
Department of Molecular Biology and Skaggs Institute
for Chemical Biology, The Scripps Research Institute, La Jolla,
CA, USA

for deriving assignments from such NMR data, since computer-based procedures allow an objective treatment of the data and enable simultaneous assessment of large quantities of data. In practice, however, the inevitable presence of spectral artefacts, absence of some expected signals, and spectral overlap impose substantial obstacles for automated resonance assignment routines. Therefore, notwithstanding a large amount of excellent work toward full automation (Atreya et al. 2002; Bartels et al. 1995, 1996; Billeter et al. 1988; Buchler et al. 1997; Coggins and Zhou 2003; Eghbalnia et al. 2005; Gronwald et al. 1998; Güntert et al. 2000; Hare and Prestegard 1994; Hyberts and Wagner 2003; Kraulis 1994; Leutner et al. 1998; Lin et al. 2005; Lukin et al. 1997; Olson and Markley 1994; Wand and Nelson 1991; Zimmerman et al. 1997), nearly all protein structure determinations published so far have used either manual approaches or computer-assisted assignment techniques in a semi-automated, interactive fashion.

In an automated approach to NMR assignment of proteins, an exhaustive search algorithm could map the NMR-identified peptide segments to their most probable positions in the primary structure. However, the inevitable presence of spectral artefacts and spectral overlap in the experimental data induce ambiguity and uncertainties into the sequential as well as the sequence-specific information. As a consequence, the cpu-time needed for an exhaustive search of the corresponding configuration space is exponentially growing with increasing protein size. This “combinatorial explosion” calls for the development of highly sophisticated assignment algorithms, since purely deterministic approaches, such as exhaustive search algorithms, are applicable only for systems with experimental input data of optimal quality. Otherwise, techniques must be employed that enable the algorithms to identify and avoid irrelevant regions of the configuration space. The field of combinatorial optimization in information technology works with algorithms that are in principle applicable to the resonance assignment problem. The program MATCH (memetic algorithm and combinatorial optimization heuristics) makes use of a memetic algorithm that enables combined use of local and global optimization heuristics. In the present implementation it is particularly efficient for obtaining sequence-specific NMR assignments for proteins with an input of APSY-NMR data (Hiller et al. 2005; Fiorito et al. 2006).

Assignment strategy and algorithms

Graph presentation of the NMR assignment problem in proteins

As a starting point for the present treatment we present the sequence-specific resonance assignment problem by two types of graphs, with the “template graph” describing the expected data, and “measured graphs” representing the experimental data obtained with a particular NMR experiment (Fig. 1). As an illustration of the use of these graphs, let us assume that extensive spin-system identification was achieved prior to the optimization, which then enables using the strategy for assignment with homonuclear ^{1}H NMR data (Wüthrich 1986). Sequential assignment is then analogous to the identification of correct links between the amino acid types corresponding to the spin systems, and sequence-specific assignment is analogous to the mapping of the measured graphs onto the primary structure represented by the template graph.

In the graph presentation of Fig. 1, the NMR assignment problem for proteins is reducible to the well-known subgraph isomorphism problem (Ullman 1976; Garey and Johnson 1979), which in turn is known to be NP-complete. This means that the NMR assignment problem can be solved by an algorithm for which the computational time is polynomial in the size of the input (NP), and a fast algorithm capable of solving this problem can be used to efficiently solve all other NP problems.

Local and global optimization algorithms

In general, the algorithms for solving the resonance assignment problem may be classified as either local and global optimization. Local optimization algorithms refine a preliminary solution by screening the adjacent configuration space in search of information on the best candidate solution. For the resonance assignment problem this is equivalent to inducing local changes to a preliminary global assignment. Local optimization can work in a highly deterministic fashion, following a concrete optimization strategy. The benefit of local optimization is high efficiency resulting from the assumption that the underlying data do not contain information that is incompatible with the rationale used by the algorithm. Local optimization thus gains efficiency at the expense of robustness. Global optimization algorithms solve combinatorial problems by optimizing all problem parameters independently. Usually they are implemented in a population-based fashion, so that multiple candidate solutions located in different regions of the configuration space are simultaneously optimized. A certain degree of randomness may be involved, analogous to mutation in biological evolution (e.g., genetic algorithms). The deteriorating influence of misleading experimental input data is thus muted, and the risk of getting trapped in local minima is greatly reduced. Overall, a population-based global optimization approach has high robustness on account of a loss of efficiency due to the fact that numerous candidate solutions have to be managed concurrently.

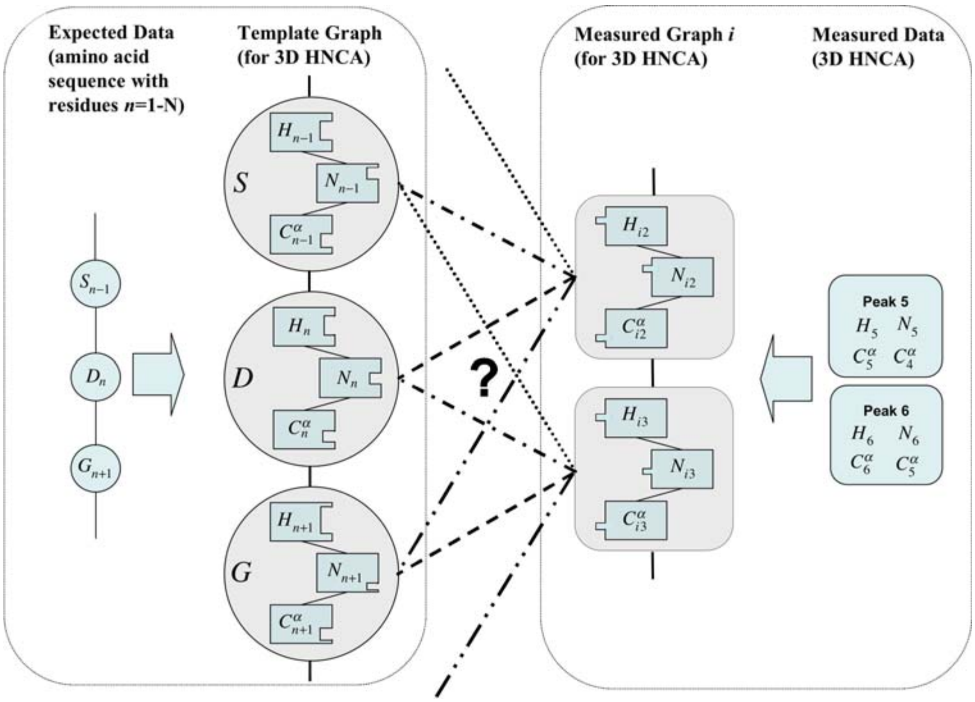

Fig. 1 Representation of protein sequence-specific resonance assignment by two types of graphs describing expected and observed data, respectively. The "template graph" of the expected data extends over the entire amino acid sequence. It is in this illustration represented by a tripeptide segment –S–D–G– from the amino acid sequence and includes knowledge about the magnetization transfer pathways in the 3D HNCA NMR experiment used. The widths of the sockets for each atom type represent the expected chemical shift ranges, as obtained from the BMRB data bank. The measured graphs, *i*, *j*, *k*, … are typically short compared to the template graph, but could in principle have any length up to that of the template graph. In this illustration, the measured graph *i* consists of a dipeptide segment of residues *i*2 and *i*3 (arbitrary numbering), and it contains the experimental NMR information from the two HNCA cross peaks on the extreme right (arbitrarily numbered 5 and 6). The chemical shifts correlated by the HNCA experiment are grouped together in the measured graph *i*, and for each atom they are represented by a plug. An assignment for the measured graph *i* is found if all its plugs fit within the sockets of a segment of equal length in the template graph. The same kind of assignment fit is searched for all measured graphs *i*, *j*, *k*, …

Memetic algorithms

Global and local optimization strategies are complementary in that local algorithms are able to optimize unambiguous regions of configuration space and arrive at correct partial solutions to the problem even when facing highly ambiguous input data, whereas in similar situations a global optimization algorithm may reject correct partial solutions in favor of an apparent global solution. A memetic algorithm is the logical attempt to merge both approaches (Moscato 1989; Corne et al. 1999; Hart et al. 2005; Ong et al. 2007), since it contains a local optimization routine embedded in an evolutionary algorithm. The evolutionary algorithm is meant to explore the overall problem space, while the local search heuristic refines discrete areas of this space. By employing the memetic approach, MATCH is able to exploit the benefits of both, local and global search heuristics, with local optimization efficiently tracing correct partial solutions inside a genetic environment that preserves robustness.

Using MATCH with APSY-NMR input data

The NMR method APSY (*Automated Projection Spectroscopy*) (Hiller et al. 2005) enables the automatic generation of high-dimensional heteronuclear correlation peak lists from analysis of a suitably selected group of experimental 2D projections of the higher-dimensional experiment. Thereby the use of high dimensions enables a significant reduction of the number of spectra needed for the resonance assignment. A further important merit of APSY spectroscopy is the determination of highly precise correlation peak chemical shifts (Fiorito et al. 2006), which is a key asset for fully automated sequence-specific resonance assignment. APSY-NMR data have previously been used as input for the automatic assignment algorithm GARANT (Bartels et al. 1996), yielding essentially complete backbone assignments of globular (Fiorito et al. 2006) and unfolded (Hiller et al. 2007) proteins. In its present implementation, MATCH has been optimized for high efficiency and reliability of automatic backbone NMR assignment of proteins when using input from APSY-NMR experiments.

Methods

The flow diagram of the new memetic algorithm MATCH for automated sequence-specific backbone resonance assignment provides a survey of 10 individual modules, which are grouped into two main building blocks, initialization and optimization (outlined with shadowed boxes in Fig. 2). Initialization includes the four modules [1] to [4] needed to load all the necessary input data, to consolidate the experimental NMR data, to generate an initial set of measured graphs (Fig. 1), and to calibrate intrinsic MATCH control parameters. The result of the initialization process represents the input for the first cycle of optimization, which includes the elements [5] to [10] (Fig. 2). Each optimization cycle starts with the creation of an initial population of individuals. This is followed by multiple evolutionary cycles, each consisting of local optimization [6] and a global "cross-over" [8], where new individuals are created and low-scoring individuals are eliminated. Within each evolutionary cycle, the configuration space is reduced whenever possible [7] and, if necessary, the threshold for the assignment of a generic spin-system to a specific-sequence position is decreased [9]. The result of each round of optimization, which typically includes a large number of evolutionary cycles, is stored as a new population of "elite individuals" [10], and thereafter a next round of optimization is started with the creation of a new initial population. In the following, the individual modules [1]–[10] are described in the order that they appear in Fig. 2.

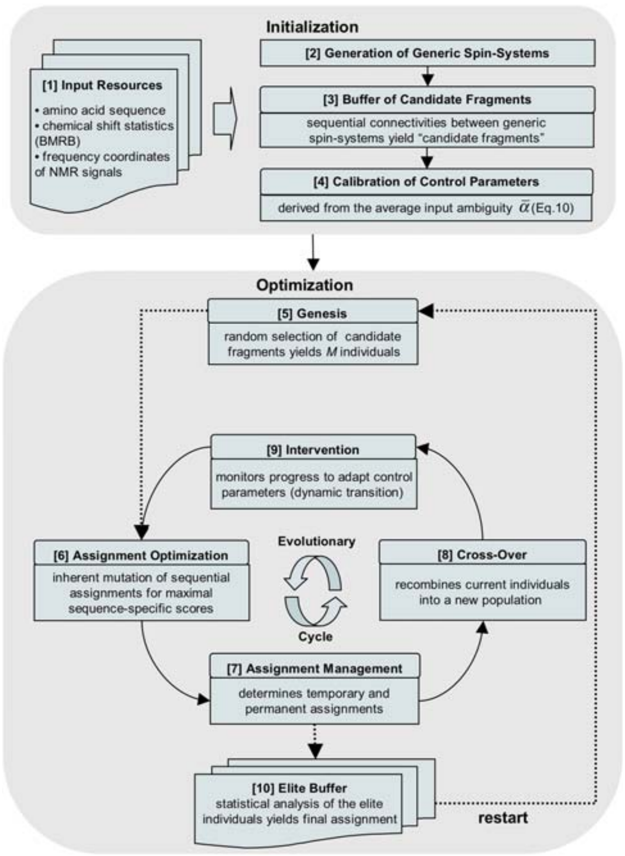
Initialization
[1] Input Resources
• amino acid sequence
• chemical shift statistics (BMRB)
• frequency coordinates of NMR signals
[2] Generation of Generic Spin-Systems
[3] Buffer of Candidate Fragments
sequential connectivities between generic spin-systems yield "candidate fragments"
[4] Calibration of Control Parameters
derived from the average input ambiguity $\bar{\alpha}$ (Eq.10)
Optimization
[5] Genesis
random selection of candidate fragments yields M individuals
[9] Intervention
monitors progress to adapt control parameters (dynamic transition)
[6] Assignment Optimization
inherent mutation of sequential assignments for maximal sequence-specific scores
Evolutionary
Cycle
[8] Cross-Over
recombines current individuals into a new population
[7] Assignment Management
determines temporary and permanent assignments
[10] Elite Buffer
statistical analysis of the elite individuals yields final assignment
restart

Fig. 2 Flow diagram of the MATCH algorithm (see text). In the optimization block, dotted arrows indicate entrance and exit pathways into and out of the evolutionary cycle, and solid arrows connect the elements within the evolutionary cycle

[1] Input resources

This module includes listings of the amino acid sequence of the protein studied, a statistical analysis of chemical shift values of proteins contained in the BioMagResBank, and the experimental input data in the form of the frequency coordinates of the NMR signals.

[2] Generation of generic spin-systems

Here, the input listings of the frequency coordinates of the NMR signals are consolidated and transformed into a single set of "generic spin-systems", G, containing all available intra- and inter-residual chemical shifts for a given spin-system,

$$G = \{g_k : k = 1, .., N\} \tag{1}$$

where N denotes the number of amino acid residues in the sequence of the protein. A generic spin-system, g_k, is designed to be composed of maximally 6 intra- and 12 inter-residual frequencies (Fig. 3):

$$g_k \equiv \left(\Omega^k_{i-1}, \Omega^k_i, \Omega^k_{i+1}\right) \tag{2}$$

$$\begin{aligned}\Omega^k_i \equiv \Big(&\omega^k(^1H^N_i), \omega^k(^{15}N_i), \omega^k(^1H^\alpha_i), \omega^k(^{13}C^\alpha_i),\\ &\omega^k(^{13}C^\beta_i), \omega^k(^{13}C'_i)\Big) \equiv \left(\omega^k(a) : a \in A\right).\end{aligned} \tag{3}$$

A is a set of atoms a that includes all backbone atoms and C^β, and the index i denotes the unknown sequence position. Each systemg_k may exist in multiple states, which enables to cope with unresolved ambiguities during the consolidation process, since each frequency dimension of g_k is then allowed to be degenerate. Different values of the individual variables $\omega^k(a)$ in Eq. 3 may then represent sets of possible values for the resonance frequency of the atom a.

[3] Buffer of candidate fragments

A graph exploration routine identifies all possible sequential connectivities between generic spin-systems up to a user-given maximal length of the resulting fragments, which is imposed by computer-technical considerations. To

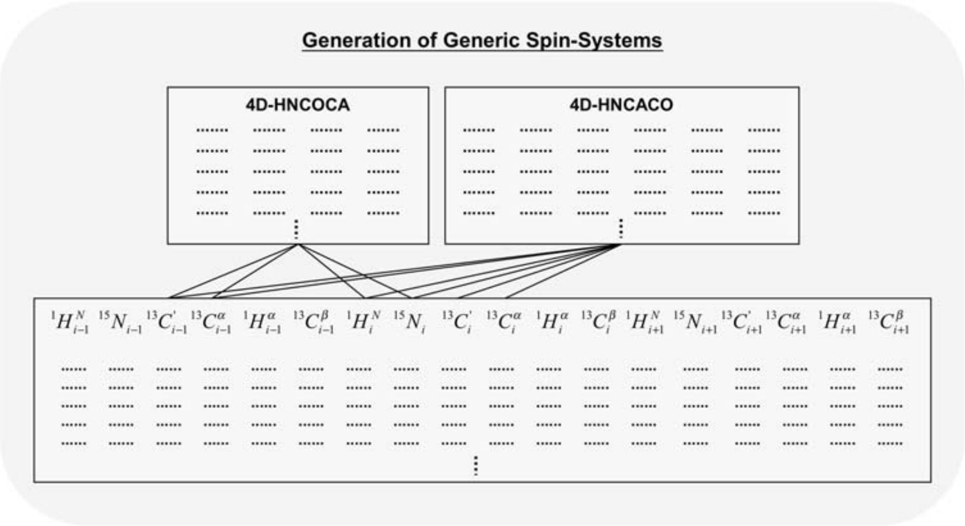

Fig. 3 Consolidation of a listing of 4D-APSY-HNCOCA and a 4D-APSY-HNCACO data into a single list of generic spin-systems. The dotted lines represent entry locations for chemical shifts, where each row is a complete set for the given NMR experiment, and for the generic spin system, respectively

each generic spin-system $g_k \in G$, a set of sequentially connected fragments is then associated,

$$s(g_k) = \{\overline{g_k g_{l_1}}, \overline{g_k g_{l_1} g_{l_2}}, \ldots, \overline{g_k g_{l_1} \cdots g_{l_n}} : g_k, g_{l_1}, g_{l_2}, \ldots, g_{l_n} \in G\}, \tag{4}$$

with maximal fragment length

$$l^{\max}_{s(g_k)} = \mathrm{n} + 1. \tag{5}$$

Thereby, n is the number of generic spin systems used to generate the longest fragment. To ensure quick access to sequential information during the optimization routine, we then create a "buffer of candidate fragments",

$$S \equiv \{s(g_k) : g_k \in G\}, \tag{6}$$

which contains all sequentially connected fragments of generic spin-systems.

For the identification of sequential connectivities, MATCH employs a scoring function that consists essentially of a series of box potentials. There is a sequential connectivity, $\overline{g_k g_l}$, between two generic spin-systems, g_k and g_l, if a set of inter-residual frequencies associated with g_k, $\{\omega^k(a) : a \in A'\}$, match their intra-residual counterparts in g_l, $\{\omega^l(a) : a \in A'\}$, within a user-specified tolerance window, $\{\Delta\omega(a) : a \in A'\}$,

$$\prod_{a \in A} \Theta\left(\Delta\omega(a) - \left|\omega^k(a) - \omega^l(a)\right|\right) = |A'|, \tag{7}$$

where

$$\Theta(x) = \begin{Bmatrix} 1 : x \geq 0 \\ 0 : x < 0 \end{Bmatrix} \tag{8}$$

is the Heaviside step function, and A' denotes the subset of A that is used to establish sequential connectivities.

At this point, a first reduction of the configuration space is possible. All generic spin-systems that are not sequentially compatible with any other generic spin-system, $s(g_k) = \emptyset$, are discarded from further consideration. This may, for example, include spurious spin systems derived from experimental artefacts in the NMR spectra.

[4] Calibration of control parameters

All relevant control parameters used in the optimization routine are automatically adapted to the degree of ambiguity contained in the experimental input data (Table 1). To estimate the degree of ambiguity, the graph exploration routine of the buffer of candidate fragments determines the set of all possible dipeptides represented by two sequentially connected generic spin-systems,

Table 1 Dependence of the control parameters of MATCH on the ambiguity, $\bar{\alpha}$, of the experimental input data (Eq. 10)

	Fast	Moderate	Slow	Very slow
	$\bar{\alpha} \in [1.0, 1.5[$	$\bar{\alpha} \in [1.5, 2.0[$	$\bar{\alpha} \in [2.0, 2.5[$	$\bar{\alpha} \in [2.5, \infty]$
F^a_{cut}	0.75	0.75	0.8	0.85
$M^a_{\min}$(Eq. 18)	0.9	0.85	0.75	0.6
F^c_{cut}(Eq. 17)	0.1	0.1	0.1	0.1

Since the efficiency of the evolutionary algorithm scales with the size of the population used, which increases with $\bar{\alpha}$ (Eq. 11), the four columns correlate with fast, moderate, slow and very slow convergence of MATCH

$$D \equiv \{\overline{g_k g_l} : g_k, g_l \in G\} \subseteq S \tag{9}$$

within the user-given chemical shift tolerance window of Eq. 7. The average ambiguity per generic spin-system is then calculated by Eq. 10,

$$\bar{\alpha} = \frac{|D|}{|G|}, \tag{10}$$

where $|D|$and $|G|$ denote the number of elements of the sets D and G, respectively. The Table 1 shows how the control parameters of MATCH are adjusted to the value of $\bar{\alpha}$, which is calculated with Eq. 10 by counting each dipeptide once, so that values of $\bar{\alpha} > 1.0$ indicate that there are degenerate connectivities.

[5] Genesis

The optimization, which is the core of the memetic algorithm, starts with the “Genesis”, where an initial population of individuals is created, each representing a projection of a set of measured graphs onto the template graph (Fig. 4). The size of the population, M, influences directly the optimization process, since small populations enable fast convergence at the expense of robustness, while large populations provide robustness but perform inefficiently. Based on numerical simulations, we use the empirical formula (11) to adjust the population size to a given value of the input ambiguity $\bar{\alpha}$ Eq. 10,

$$M \equiv \begin{cases} 50 \cdot e^{1.5 \cdot (\bar{\alpha}-1.0)} & 1.0 \le \bar{\alpha} < 1.5 \\ 75 \cdot e^{2.0 \cdot (\bar{\alpha}-1.5)} & 1.5 \le \bar{\alpha} < 2.0 \\ 125 \cdot e^{2.0 \cdot (\bar{\alpha}-2.0)} & 2.0 \le \bar{\alpha} < 2.5 \\ \min(500, 200 \cdot e^{2.0 \cdot (\bar{\alpha}-2.5)}) & \bar{\alpha} \ge 2.5 \end{cases} \tag{11}$$

The scheme used to generate new individuals by mapping fragments from the buffer [3] onto the amino acid sequence is governed by the current fragment length, l^c, that is initially set to the maximal allowed fragment length given in Eq. 5,

$$l^c = MAX\left\{ l^{\max}_{s(g_m)} : g_m \in G \right\}. \tag{12}$$

A generic spin-system, $g_k \in G$, associated with a fragment of length

$$l^{\max}_{s(g_k)} = l^c \tag{13}$$

is then randomly selected from the buffer [3]. The “sequence space” represented by the template graph is then screened so as to map this candidate fragment onto the position with the highest value of the “sequence-specific scoring function” (see below). Thereby, if g_k is associated with multiple fragments of length l^c, one of these fragments is randomly chosen and mapped to a sequence position. All generic spin-systems present in the mapped fragment are then excluded from further use, the matched residues are removed from the sequence space, and the process restarts by randomly choosing a next generic spin-system. The fragment length, l^c, is reduced when either all

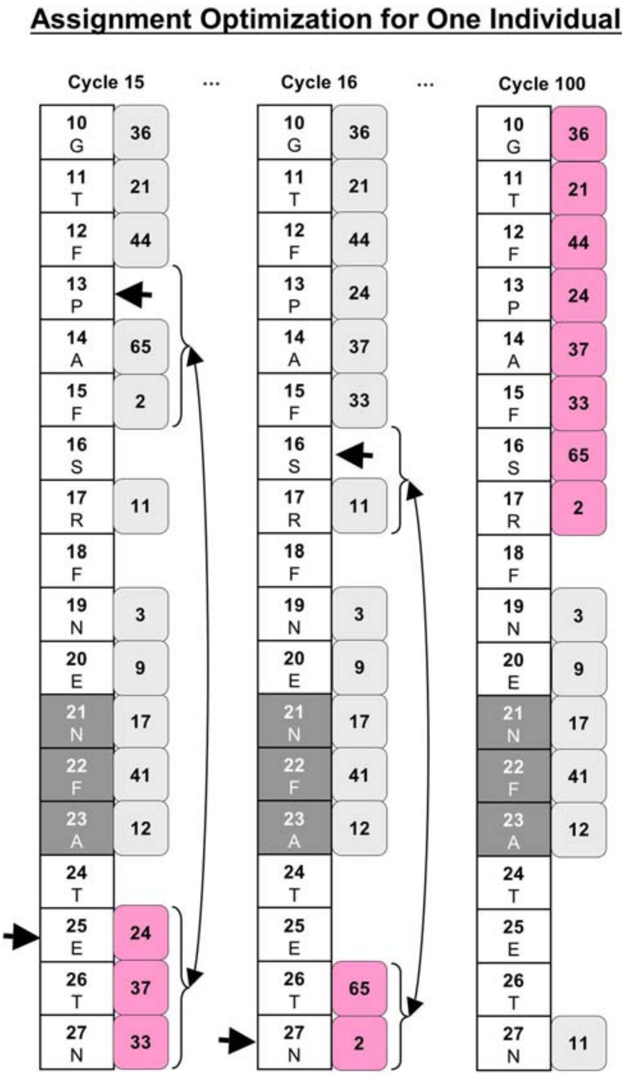

Fig. 4 Assignment optimization [6]. In the template graph (Fig. 1), the sub-space of all sequence positions to which fragments have been temporarily or permanently assigned, A^S, is marked by dark grey boxes, and the sub-space of all sequence positions to which no generic spin-systems have been assigned either temporarily or permanently, T^S, is marked by white boxes. Measured graphs of variable lengths (Fig. 1), the “candidate fragments”, are represented by light grey and pink boxes. An assignment is initiated by random selection of an element of T^S (left arrow to 25 in cycle 15) along with one of the candidate fragments that was tentatively placed along the sequence during genesis [5] (pink boxes). Next, a target position in $T^S$is randomly chosen (right arrow to 13 in cycle 15). If at least one edge of the candidate fragment sequentially matches a generic spin-system at or adjacent to the selected target position (Fig. 1), a swap is performed (long vertical double-headed arrow). The process restarts by selecting a new sequence position (arrow to 27 in cycle 16). After multiple additional cycles, the initially selected 3-residue candidate fragment could be enlarged and sequence-specifically assigned, since it satisfies Eq. 18 (pink boxes in cycle 100)

fragments of the current length have been assigned, or if no coherent sequence space is left unoccupied. This iterative scheme is continued until all generic spin-systems have been processed.

The sequence-specific scoring function indicates the probability for a given fragment of sequentially connected generic spin-systems to be compatible with a specific position in the amino acid sequence of the protein. To this end, MATCH employs a χ^2-significance test for determining the goodness-of-fit of an observed distribution to a statistical distribution of chemical shift values in proteins taken from the BioMagResBank. If ω_i are k independent, normally distributed random variables (here chemical shifts) with mean valuesμ_i and variances σ_i, then the random variable Z-score,

$$Z = \sum_{i=1}^{k} \left(\frac{\omega_i - \mu_i}{\sigma_i} \right)^2 \geq 0, \tag{14}$$

is distributed according to the χ^2-distribution. In the probability density distribution of χ^2, the shape parameter, k, specifies the number of degrees of freedom:

$$\chi^2(Z,k) = \frac{(1/2)^{k/2}}{\Gamma(k/2)} Z^{k/2-1} e^{-Z/2}; \quad Z \geq 0, \tag{15}$$

where Γ denotes the Gamma function. The cumulative χ^2-distribution function is defined as

$$F(Z,k) = \frac{\gamma(k/2, Z/2)}{\Gamma(k/2)}; \quad Z \geq 0, \tag{16}$$

where γ is the incomplete Gamma function. Thus, $F(Z,k)$ represents the probability that a given distribution of random variables Z (here a given set of chemical shifts) matches a reference distribution (here the expected chemical shifts at specific positions in the amino acid sequence of the protein).

[6] Assignment optimization

Assignment optimization is a local optimization step which bears on the length and composition of the candidate fragments from [3] and on their sequence-specific assignment, and is performed independently for each individual (Figs. 2 and 4). At any stage of the iteration, two subspaces of the sequence space S are determined, $A^S$and T^S, where S is the amino acid sequence of the protein represented by the template graph in Fig. 1. A^S is the sub-space of all sequence positions to which candidate fragments (represented by measured graphs in Fig. 1) have been temporarily or permanently assigned, and T^S is the subspace of all sequence positions without such temporary nor permanent assignments. Optimization is initiated by random selection of a position in T^S along with a candidate fragment. Next, a sequence position in $T^S$is randomly selected and the candidate fragment is checked for compatibility with the spin-systems at and adjacent to the new position. In cases where there is a sequential match for at least one edge of the candidate fragment, a swap of candidate fragments is performed. Thereby, each candidate fragment is given a fixed maximal number of attempts to find a compatible sequence location. If this contingent of attempts is used up, the procedure is started with a new candidate fragment. For each candidate fragment the sequence-specific score (Eq. 16) is eventually evaluated for all possible (i.e., so far unassigned) sequence positions. If the sequence-specific score, $F(Z,k)^c$, satisfies Eq. 17,

$$F(Z,k)^c \geq F^c_{\text{cut}}, \tag{17}$$

where $F^c_{\text{cut}}$is initially set to a small value, so as to avoid a dominant impact of the local optimization (Table 1), then an assignment is made. If Eq. 18 is satisfied for multiple sequence positions, then one of these sequence positions is randomly selected for an assignment. Once the target position is determined, the candidate fragment remains unchanged until the end of the optimization for the individual considered. The concept of "inherent mutation" used in this approach can thus recombine current sequential assignments by searching for arrangements of all candidate fragments that are compatible with the amino acid sequence.

[7] Assignment management

Whenever a local optimization step [6] in the evolutionary cycle has been completed, the resulting assignments of candidate fragments are newly evaluated. "Temporary assignments" are associated with individuals of the population M (Eq. 11), whereas "permanent assignments" are associated with all individuals within M (Fig. 5). First, all previously stored temporary sequence-specific assignments are removed from all individuals. Second, each individual is reassessed: if the sequence-specific score of a candidate fragment, $F(Z,k)^c$ (Eq. 16), is above a predetermined assignment threshold, F^a_{cut} (Table 1 and Eq. 17), a temporary sequence-specific assignment is stored. The thus assigned fragment will be excluded from the subsequent local optimization step, so that the problem space is temporarily reduced and the efficiency of the subsequent process is improved.

Simultaneously, a cross-check throughout the entire population M (Eq. 11) is performed. If the frequency with which a fragment is mapped to a specific sequence position, M^a, satisfies Eq. 18,

$$M^a \geq M^a_{\text{min}}, \tag{18}$$

a permanent sequence-specific assignment is stored (for values of M^a_{min} see Table 1). The fragment concerned is removed from the problem space and will remain mapped

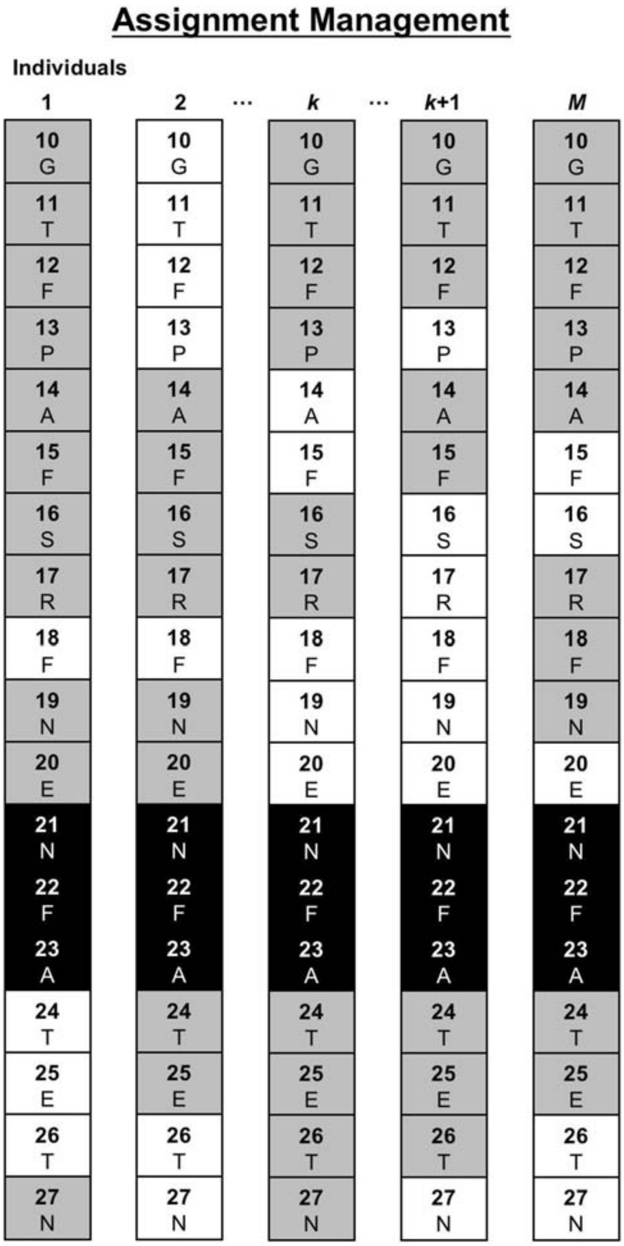

Fig. 5 Assignment management [7]. Illustration of temporary and permanent assignments made by MATCH throughout a population of individuals, *M* (Eq. 11). Black, permanent assignments; grey, temporary assignments; white, no assignments

to the permanent sequence position throughout the remainder of the optimization process (Fig. 5), thus increasing the efficiency of the process.

[8] Cross-over

"Cross-over" is the key module of the evolutionary cycle. It identifies the most promising individuals based on their sequential and sequence-specific information. In addition to commonly used routines, MATCH supports population size control and recombination of more than two individuals. This makes sense when dealing with highly ambiguous input data, since this approach favors the probability of finding high-scoring individuals as well as apparently correct fragments.

Initially, the parental population of individuals available after the assignment management [7] is ranked according to their sequence-specific scores (Eq. 16), and the sequential and sequence-specific information of the parental individuals is assigned to a repository ("gene pool"). A fraction of the best-scoring individuals is then selected for the cross-over, where the content of the gene pool is transformed into a new population of individuals. This is achieved by sorting the gene pool according to fragment lengths and sequence-specific scores. It is sensible to prefer long fragments, because their scores are a more reliable indication for the correctness of the assignment. Analogous to the procedure applied in the "Genesis" [5], the maximal fragment length is defined (Eq. 12). The corresponding fragments are mapped to their inherited sequence positions on the new individual. As soon as all fragments of the current length have been assigned, the maximal length is decreased to the next possible value, which again corresponds to the procedure [5]. The generation of the new individual is complete as soon as all available generic spin-systems have thus been processed.

[9] Intervention

In order to monitor the progress of the optimization process and to adapt the intrinsic control parameters, MATCH determines the population homogeneity, H, after each step in the evolutionary cycle, according to

$$H \equiv \frac{T - N}{|G|N} \qquad (19)$$

where T, N and $|G|$denote the total number of different sequence-specific resonance assignments of all generic spin systems throughout the whole population, the number of amino acids in the amino acid sequence, and the number of generic spin-systems, respectively. The homogeneity shows a cyclic behavior during the optimization process. The local optimization reduces the level of homogeneity due to inherent mutation and the local exploration of the configuration space. During the cross-over the homogeneity is again increased. If the homogeneity shows no overall upward trend during a specified number of optimization cycles, the control parameters may have been set too restrictively, so that they impede convergence. In this case, MATCH intervenes and adapts the control parameters.

[10] Elite buffer

The optimization for a given individual is completed when either all generic spin-systems are permanently assigned, or the total sequence-specific scores of all individuals are equal. The second criterion would enable the determination of a final result even if incompatible apparent generic spin-systems, which might be due to noise peaks or artefacts, had not been eliminated in the course of the optimization (see above). These are then not included in the sequence-specific scoring, and are therefore readily identified when all individuals have reached equal scores. The sequence-specific assignment thus obtained is added to a new population of individuals stored in the "elite buffer" ([10]), and the optimization restarts until a predetermined number of elite individuals have been generated. Sequence-specific resonance assignments which occur in more than 50% of the elite individuals are accepted as being correct and are printed out together with their sequence-specific scores. Deleted and oscillating generic spin-systems are listed in a supplementary output.

Results and discussion

Sequence-specific resonance assignment of the protein TM1290

A list of NMR frequency positions for the 115-residue hypothetical protein TM1290 from *Thermotoga maritima*, which had been automatically generated with the GAPRO algorithm (Hiller et al. 2005; Fiorito et al. 2006) from a 6D-APSY-seq-HNCOCANH spectrum, was used as input for the MATCH algorithm. The list contained 98 6D-correlations. MATCH was instructed to generate 30 elite individuals, i.e., to perform 30 independent sequence-specific resonance assignments. Despite the fact that APSY-NMR yields highly precise peak positions (Fiorito et al. 2006), the tolerance windows used for sequential matching were set to rather large values, i.e., $\Delta\omega(^1H^N) = 0.05$ ppm and $\Delta\omega(^{15}N) = 0.4$ ppm. On average, 9.3 evolutionary cycles were performed per optimization run, with a calculation time for each individual sequence-specific resonance assignment in the range of 10–15 s.

The result obtained using automatic assignment with MATCH is identical to the previous sequence-specific assignment obtained with an interactive approach (Etezady-Esfarjani et al. 2003). The presentation of the data in Fig. 6 shows for all residues for which an APSY-NMR correlation could be observed (see caption to Fig. 6), that they were correctly assigned. For all but five of the assigned residues, all 30 elite individuals yielded identical assignments, and for the remaining five residues the correct assignment was obtained from 29 of the 30 elite individuals. These results are far above the requirement (see section [10] above) that at least 50% of the elite individuals must yield an identical result for the sequence-specific assignment to be accepted as a valid solution.

Similar results to those for TM1290 were obtained with several other proteins for which high-quality 4D, 5D or 6D APSY-NMR data sets could be recorded. These proteins were selected as targets in an on-going structural genomics project and represent different molecular sizes and different secondary structure types. For most of these proteins, the input data consisted of a combination of two or three 4D and 5D APSY-NMR experiments, whereby these combinations were selected so that they had a similar information content as the presently used 6D APSY-NMR data set of TM1290. Although lower-dimensional spectra quite naturally contain more extensive peak overlap than a 6D experiment, MATCH performed equally well with an input consisting of a combination of lower-dimensional peak lists (B. Pedrini, private communication). Based on the experience gained so far in terms of robustness and efficiency of the MATCH algorithm, we will broaden the application range by including conventional triple-resonance NMR experiments as input data.

Robustness tests

The automatic MATCH assignment of the 115-residue hypothetical protein TM1290 (Fig. 6) was based on a complete, artifact-free input of 6D correlations. In order to assess the robustness of the MATCH algorithm when faced with less complete or less precise input data, we deteriorated this peak list which had been automatically generated from a 6D-APSY-seq-HNCOCANH spectrum and contained 98 NMR signals. Thereby, instead of simply adding random noise to the experimental data set, variant data sets were generated by elimination of discrete sets of peaks, controlled variation of chemical shifts, and recombination of individual ones of the sets of six experimental chemical shifts per peak into new, artefactual 6D correlation peaks. To account for statistical outliers at a given extent of deterioration, multiple MATCH calculations were performed, and the performance of the program was evaluated from comparison of the percentage of incorrectly assigned peaks among the assignable peaks, Δ, and from the average number of evolutionary cycles (Fig. 2) needed to obtain convergence to the elite individuals, *C*.

Elimination of peaks is equivalent to a decrease of the average length of some candidate fragments and thus challenges the sequence-specific scoring (Fig. 7). By experience, if the average fragment length is below 6, then the sequence-specific scoring gradually loses reliability and the quality of the sequence-specific assignment decreases. In the test of Fig. 7, we observe a virtually flawless performance if up to 10% of the peaks are missing, and the average number of incorrectly assigned peaks were below 8% even if 20% of the peaks were removed. Thereby, the number of evolutionary cycles needed to achieve

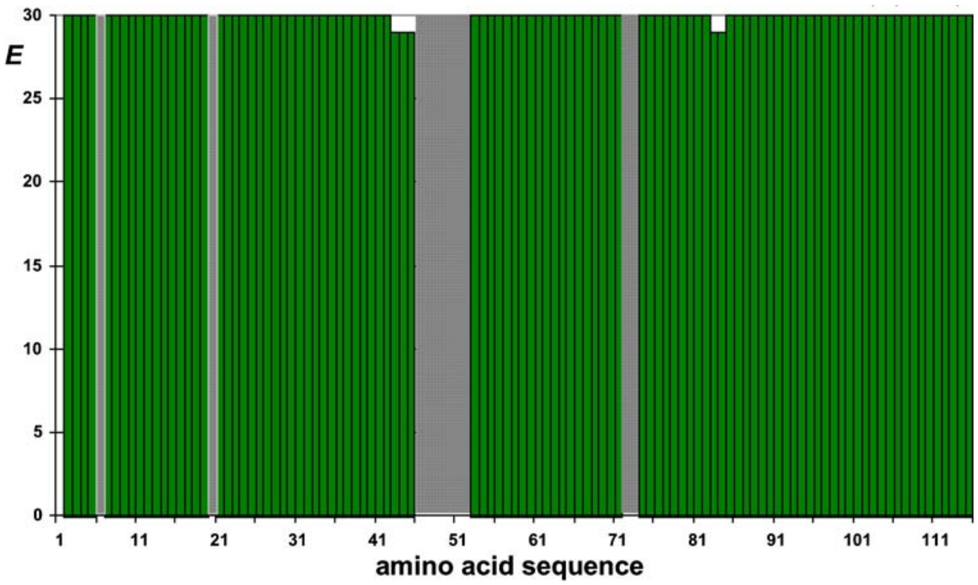

Fig. 6 Assignment statistics for TM1290. The input peak list of the protein had been automatically generated from a 6D-APSY-seq-HNCOCANH spectrum using the GAPRO algorithm with standard parameter set. MATCH had been instructed to generate 30 elite individuals, *E* (Fig. 2), which are represented along the vertical axis. For each sequence position along the horizontal axis, a green column represents the number of identical, correct sequence-specific assignments. The grey areas indicate that there are no peaks present in the input peak list which could be mapped to the given sequence position, either because the positions are occupied by prolyl residues, or because the NMR signals were broadened beyond detection by slow dynamic processes (Etezady-Esfarjani et al. 2003)

convergence increased at an exponential rate with the extent of elimination (Fig. 7). This behavior of the algorithm results from the use of dynamic transition. MATCH absorbs the increasing ambiguity of the sequence-specific assignments in the local assignment optimization by frequent switching to global, population-based optimization.

Variation of the input chemical shifts (Fig. 8) challenges again the sequence-specific scoring. In this computer experiment, all input peaks were simultaneously manipulated by random variation of the frequency coordinates within a predetermined interval around the experimental values. The range of the interval is defined by a given number of standard deviations for each atom involved in a correlation peak, which were taken from the BioMagResBank. MATCH performed highly reliably and with good efficiency as long as the chemical shift variation was below one standard deviation (Fig. 8). Again, dynamic transition enabled MATCH to cope with the decreased accuracy of the sequence-specific scoring. For chemical shift variations above one standard deviation, the performance of MATCH deteriorates dramatically, emphasizing the importance of precise frequency measurements. This behavior reflects an important feature of the χ^2-distribution (Eq. 15): with exponentially increasing *Z*-scores, the cumulative χ^2-distribution (Eq. 16) rapidly goes toward 0 if the deviation from the expected values exceeds one standard deviation.

In the experiment of Fig. 9, elimination of correct input signals (Fig. 7) is combined with the admixture of spurious signals, which is achieved by recombination of the sets of six experimental chemical shifts of a predetermined fraction of all input peaks into new, artefactual 6D correlation peaks. MATCH performed reliably and efficiently as long as the extent of recombination was below 10%. Comparison with Fig. 7 shows further that the introduction of up to 10% of spurious peaks has no measurable effect on the outcome of the assignment. More extensive addition of

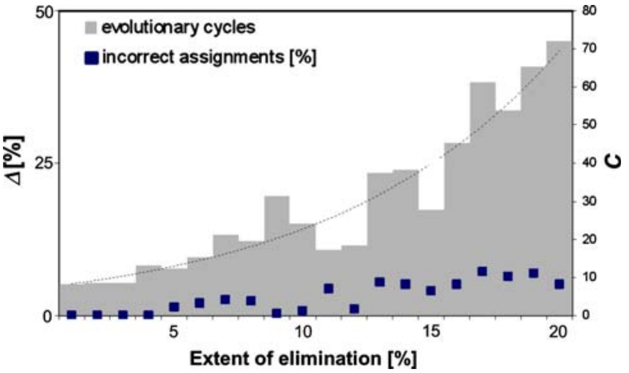

Fig. 7 Robustness test of MATCH when facing incomplete input peak lists. From the input used in Fig. 6, variable percentages of the 6D APSY-NMR peaks were deleted before the assignment process was started, whereby different selections of peaks were deleted in each of the calculations resulting in the generation of 30 elite individuals for each extent of elimination. Δ is the percentage of erroneous assignments obtained from analysis of the 30 elite individuals, and *C* is the average number of evolutionary cycles (Fig. 2) needed to achieve convergence to the elite individuals

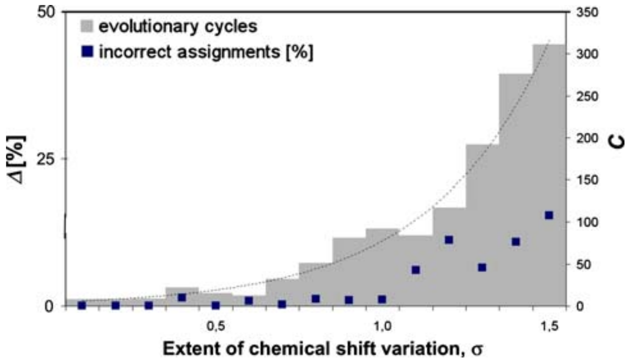

Fig. 8 Same presentation as in Fig. 7 for the results of a study on the effects of chemical shift variation in the complete input set of peaks used in Fig. 6. All the input peaks were simultaneously manipulated by random variation of all six frequency coordinates. The extent of the shift variations is given in units of the standard deviation, σ, about the experimental values. The standard deviations were taken from the BMRB

artefacts overburdens inherent mutation, because recombined peaks may be very similar to real peaks, and reduced performance of MATCH is observed when compared to the situation arising from deletions of the input data without introduction of artefacts (Fig. 7).

Overall, the data of Figs. 7–9 show that MATCH is able to cope with significantly lower quality experimental input data than those obtained from 6D APSY-NMR in Fig. 6, including the admixture of artifactual peaks, imprecise frequency positions and missing signals. Current applications of MATCH for backbone resonance assignment of a variety of proteins are in agreement with the computer simulations presented above (B. Pedrini, personal communications). This indicates that MATCH should also be applicable for automatic backbone assignment using input measured with conventional triple-resonance NMR experiments. In the present implementation of MATCH such applications would require that the format of the input peak lists correspond to the format of the GAPRO output from APSY-NMR experiments (Hiller et al. 2005).

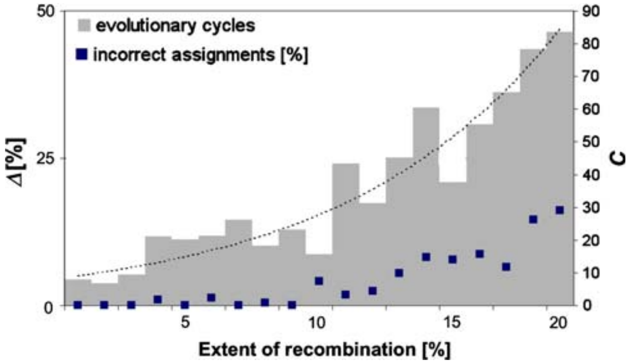

Fig. 9 Same presentation as in Fig. 7 for the results of a study on the effects of deteriorating the complete input used in Fig. 6 by recombination of the six experimental chemical shifts of a predetermined percentage of the peaks into new, artifactual 6D correlation peaks

For academic users, MATCH will be distributed free-of-charge as a module of the stand-alone ATNOS/CANDID program (Herrmann et al. 2002a, b). Download information is available under http://www.mol.biol.ethz.ch/groups/wuthrich_group/software.

Acknowledgments We thank Dr. B. Pedrini for sharing his experience with applications of MATCH for backbone resonance assignments in a variety of proteins. Financial support by the Schweizerischer Nationalfonds (project 3100-AO-113838) is gratefully acknowledged.

References

Atreya HS, Chary KVR, Govil G (2002) Automated NMR assignments of proteins for high throughput structure determination: TATAPRO II. Curr Sci 83:1372–1376

Bartels C, Xia C-H, Billeter M, Güntert P, Wüthrich K (1995) The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. J Biomol NMR 5:1–10

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

Billeter M, Basus VJ, Kuntz ID (1988) ID: a program for semi-automatic sequential resonance assignments in protein ^{1}H nuclear magnetic resonance spectra. J Magn Reson 76:400–415

Buchler NEG, Zuiderweg ERP, Wang H, Goldstein RA (1997) Protein heteronuclear NMR assignments using mean-field simulated annealing. J Magn Reson 125:34–42

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

Corne D, Dorigo M, Glover F (1999) New ideas in optimization. McGraw-Hill

Eghbalnia HR, Bahrami A, Wang L, Assadi A, Markley JL (2005) Probabilistic identification of spin systems and their assignments including coil-helix inference output (PISTACHIO). J Biomol NMR: 219–233

Etezady-Esfarjani T, Peti W, Wüthrich K (2003) NMR assignment of the conserved hypothetical protein TM1290 of *Thermotoga maritima*. J Biomol NMR 25:167–168

Fiorito F, Hiller S, Wider G, Wüthrich K (2006) Automated resonance assignment of proteins: 6D APSY-NMR. J Biomol NMR 35:27–37

Garey MR, Johnson DS (1979) Computers and intractability. A guide to the theory of NP-completeness. Freeman, New York

Gronwald W, Willard L, Jellard T, Boyko RF, Rajarathnam K, Wishart DS, Sönnichsen FD, Sykes BD (1998) CAMRA, chemical shift based computer aided protein NMR assignment. J Biomol NMR 12:395–405

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

Hare BJ, Prestegard H (1994) Application of neural networks to automated assignment of NMR spectra of proteins. J Biomol NMR 4:35–46

Hart WE, Krasnogor N, Smith JE (2005) Recent advances in memetic algorithms series: studies in fuzziness and soft computing, vol 166. Springer, Berlin and Heidelberg

Herrmann T, Güntert P, Wüthrich K (2002a) Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. J Biomol NMR 24:171–189

Herrmann T, Güntert P, Wüthrich K (2002b) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J Mol Biol 319:209–227

Hiller S, Fiorito F, Wüthrich K, Wider G (2005) Automated projection spectroscopy (APSY). Proc Natl Acad Sci USA 102:10876–10881

Hiller S, Wasmer W, Wider G, Wüthrich K (2007) Sequence-specific resonance assignment of soluble nonglobular proteins by 7D APSY-NMR spectroscopy. J Am Chem Soc 129:10823–10828

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

Ikura M, Kay LE, Bax A (1990) A novel approach for sequential assignment of ^{1}H, ^{13}C, and ^{15}N spectra of larger proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Biochemistry 29:4659–4667

Kay LE, Ikura M, Tschudin R, Bax A (1990) Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. J Magn Reson 89:496–514

Kraulis PJ (1994) Protein three-dimensional structure determination and sequence-specific assignment of ^{13}C and ^{15}N-separated NOE data. A novel real space ab initio approach. J Mol Biol 243:696–718

Leutner M, Gschwind RM, Liermann J, Schwarz C, Gemmecker G, Kessler H (1998) Automated backbone assignment of labelled 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

Lukin JA, Gove AP, Talukdar SN, Ho C (1997) Automated probabilistic method for assigning backbone resonances of (13C, 15N)-labeled proteins. J Biomol NMR 9:151–166

Montelione GT, Wagner G (1989) Accurate measurements of homonuclear H^{N}-H^{α} coupling constants in polypeptides using heteronuclear 2D NMR experiments. J Am Chem Soc 111:5474–5475

Montelione GT, Wagner G (1990) Conformation independent sequential NMR connections in isotope-enriched polypeptides by 1H–13C-15N triple-resonance experiments. J Magn Reson 83:183–188

Moscato P (1989) On evolution, search, optimization, genetic algorithms and martial arts: Towards memetic algorithms. Caltech Concurrent Computation Program, C3P Report

Olson JB Jr, Markley JL (1994) Evaluation of an algorithm for the automated sequential assignment of protein backbone resonances: a demonstration of the connectivity tracing assignment tools (CONTRAST) software package. J Biomol NMR 4:385–410

Ong YS, Krasnogor N, Ishibuchi H (2007) Special Isssue on Memetic Algorithms. IEEE Trans Syst, Man, Cybernet, Part B 37(1):2–5

Ullman JD (1976) An algorithm for subgraph isomorphism. J ACM 23:31–42

Wand AJ, Nelson SJ (1991) Refinement of the main chain directed assignment strategy for the analysis of ^{1}H NMR spectra of proteins. Biophysics 59:1101–1112

Wüthrich K (1983) Sequential individual resonance assignments in the ^{1}H-NMR spectra of polypeptides and proteins. Biopolymers 22:131–138

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

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