

SideLink: Automated side-chain assignment of biopolymers from NMR data by relative-hypothesis-prioritization-based simulated logic

James E. Masse, Rochus Keller, Konstantin Pervushin *

Laboratorium für Physikalische Chemie, ETH Zurich, CH-8093, Zurich, Switzerland

Received 15 December 2005; revised 6 March 2006

Available online 24 April 2006

Abstract

Previously we published the development of AutoLink, a program to assign the backbone resonances of macromolecules. The primary limitation of this program has proven to be its inability to directly recognize spectral data, relying on the user to define peak positions in its input. Here, we introduce a new program for the assignment of side-chain resonances. Like AutoLink, this new program, called SideLink, uses Relative Hypothesis Prioritization to emulate “human” logic. To address the higher complexity of side-chain assignment problems, the RHP algorithm has itself been advanced, making it capable of processing almost any combinatorial logic problem. Additionally, SideLink directly examines spectral data, overcoming the need and limitations of prior data interpretation by users. © 2006 Elsevier Inc. All rights reserved.

Keywords: NMR spectroscopy; Relative hypothesis prioritization; RHP; Automation; Artificial intelligence

1. Introduction

The primary bottlenecks in the process of reconstruction of 3D protein structures by NMR are time consuming data acquisition and resonance assignment, since the subsequent steps to structure calculation have been automated [1–6], and are therefore not as human expert time-intensive [7]. Automation of resonance assignment has proven difficult (reviewed in [8,9], (for a list of programs found in the literature see [Supplementary Information Table 2](#))), largely due to the highly non-monotonic nature of the problem. By “non-monotonic” here we mean that no sub-component of the overall problem can be solved without considering the rest of the problem—i.e., assignment of any particular resonance to any particular nucleus precludes simultaneous assignment of the same resonance to another nucleus or simultaneous assignment of any other resonance to the same nucleus. This formulation matches so called “quadratic assignment problem,” which in turn might be defined as NP-hard [10]. When one adds the additional

considerations of ever present spectral artifacts (false positives), missing peaks (false negatives), peak overlap, noise in the data, and (for some protein forms) the appearance of multiple peaks in the data, the non-monotonic component of the problem is substantially increased.

Attempts in the past have been made to automate resonance assignment, which primarily rely on treatment of the non-monotonic problem in a monotonic manner, which can be solved in a polynomial ($O(n^3)$) time scale. These approaches reduce the ambiguity in the assignment problem either by acquiring more non-redundant data enabling constraint propagation towards solution in a bootstrap fashion (an approach dubbed best-first matching, i.e., Aurelia, [11], AutoAssign [12,13], DBPA/PGA/CPA/ASPA/NCPA/PMA [14–16], and SPSCAN [17]), use of an exhaustive search and enumeration of all connected spin systems followed by best-first mapping interactively approved by human expert (i.e., PACES [18]) or by using structural data as an additional data element (GARANT [19,20], *St2nmr* [21], NVR, [22,23], NOESY JIGSAW [24], CAMRA [25], and Tian et al. [26]). If enough information exists to allow the assignment problem to be broken down into subcomponents, each with a unique

* Corresponding author. Fax: +41 1 632 10 21.

E-mail address: kope@phys.chem.ethz.ch (K. Pervushin).

sub-solution, the problem can then be considered as largely monotonic with computationally tractable global mapping by combinatorial minimization strategies. In the most favorable cases sequence specific resonance assignment and structural NOE assignment can be done in parallel with structure calculations as is attempted by the program ABACUS [27,28] (see also CLOUDS [29,30], which does not itself assign chemical shifts, but rather calculates a proton density distribution which the user can then map a protein into to obtain resonance assignments). However, despite the reduction in ambiguity from the above approaches, there is often still enough complexity left to require at least some human assistance (humans can generally solve non-monotonic problems, though often only with a substantial effort and time commitment).

Attempts to solve NMR assignment problems in a non-monotonic formulation include genetic algorithms, neural networks, simulated annealing, iterative relaxation techniques (ALFA [31], ALPS[32], MONTE [33], GARANT [19,20], Buchler et al. [34], PASTA [35], DBPA [14–16], CAPRI [36], GANA [37], TATAPRO [38,39], RIBRA [40], and HIPS [41–43]). PISTACHIO [10] (earlier CONTRAST [44]) transforms the deterministic, combinatorial optimization problem into a search for the ground state configuration of a statistical system using combinations of chemical shifts (and possibly all available structural information) in tripeptides to reduce the search space. MARS [45] uses a hybrid strategy where the assignment is driven by best-first mapping and reevaluated by global scoring.

Recently we developed AutoLink [46], a novel program designed to automatically determine backbone resonances in macromolecules. This program relies on human logic emulation by “Relative Hypothesis Prioritization” to treat NMR data in a very “human-spectroscopist-like” way. Since the program takes a human approach, it can directly handle the problem’s non-monotonic nature. It also, therefore, requires no data other than that which a human spectroscopist would require, and thus, no novel NMR experiments are required to apply the program.

Now we have turned our focus to the next step in the resonance assignment process, that of side-chain assignment. Only a few of the programs designed to automate NMR assignment problem tackle (with a very moderate success) this more challenging task (i.e., RESCUE [47–49], DBPA/PGA/CPA/ASPA/NCPA/PMA [14–16], GARANT [19,20], and ABACUS [27]). For this purpose we have developed a new program call SideLink. Like AutoLink, SideLink approaches the problem in a human-spectroscopist-like manner. The program uses a more-sophisticated version of the relative hypothesis prioritization (RHP) algorithm, significantly expanded in order to allow it to handle the much more complex logic involved in side-chain assignment. The program can use any available input assignments the user can provide (especially those determined during the backbone assignment process), but alternatively can also function from

minimal backbone resonance assignments. SideLink can also work from a variety of spectrum types as input, but has primarily been designed to work from NOESY spectra, which are ubiquitous for structure determination by NMR. In addition to improvement of the RHP logic engine, SideLink has an additional advantage over AutoLink—it can access spectra directly, and, thus, there is no need for any prior spectral analysis by the user in the form of peak picking or defining of peak folding. Instead of working on picked peaks, SideLink relies on mathematical comparison of spectrum slices to associate resonance frequencies. Since the program does not require peak input from the user, it can function completely autonomously once the user has supplied the spectra, the backbone assignments, and control parameters.

The main focus of this article is a non-exhaustive description of the expanded RHP algorithm that drives the program and its application to the resonance assignment of side-chains of protein residues. We argue that NOESY-type spectra are often sufficient to assign all side-chain resonances in small and medium-sized proteins. Use of the program is demonstrated on two test cases, working from minimal 3D NOESY spectra. These tests show that SideLink can assign >80% of the side-chain [$^1\text{H-X}$] resonances (only residues with known backbone assignments are considered) with >95% accuracy.

2. Methods

2.1. Formalizing the side-chain assignment problem

Typically prior to assigning the side-chains, the user will already have assigned the backbone, including the amides, most of the C_{α} s, and possibly some of the C_{β} s and/or carbonyl carbons. Additionally, some inter-residue information is usually available from inter-residue cross-peaks (such as the amide $\text{HN} \rightarrow C_{\alpha-1}$ cross-peaks in HNCA spectra). In most cases, the remaining side-chain resonances must be assigned using additional spectra beyond what was required to obtain backbone assignments. This is simply due to the fact that most of the side-chain resonances are not apparent in the backbone assignment spectra.

The most commonly used spectra for side-chain assignment are the 3D HCCH-COSY [50–52], 3D HCCH-TOCSY [53,54], HCC(CO)NH-TOCSY [55,56], HCCNH-TOCSY [55,57], and the 3D ^{13}C -resolved NOESY [58,59], though other types of spectra are often used in addition [60–63]. Unfortunately, neither the HCCH-TOCSY nor the ^{13}C -resolved NOESY contain any clear correlation data with regard to the backbone amide frequencies. Thus, to use them to assign side-chains, the spectroscopist must visually inspect the spectra and logically determine which side-chain resonances belong to which residue based on comparison with the known resonances of the backbone and their cross-peaks in the backbone assignment spectra (see Fig. 1). Typically this means the user spends days comparing 1D slices from the ^{13}C -resolved spectra with 1D slices

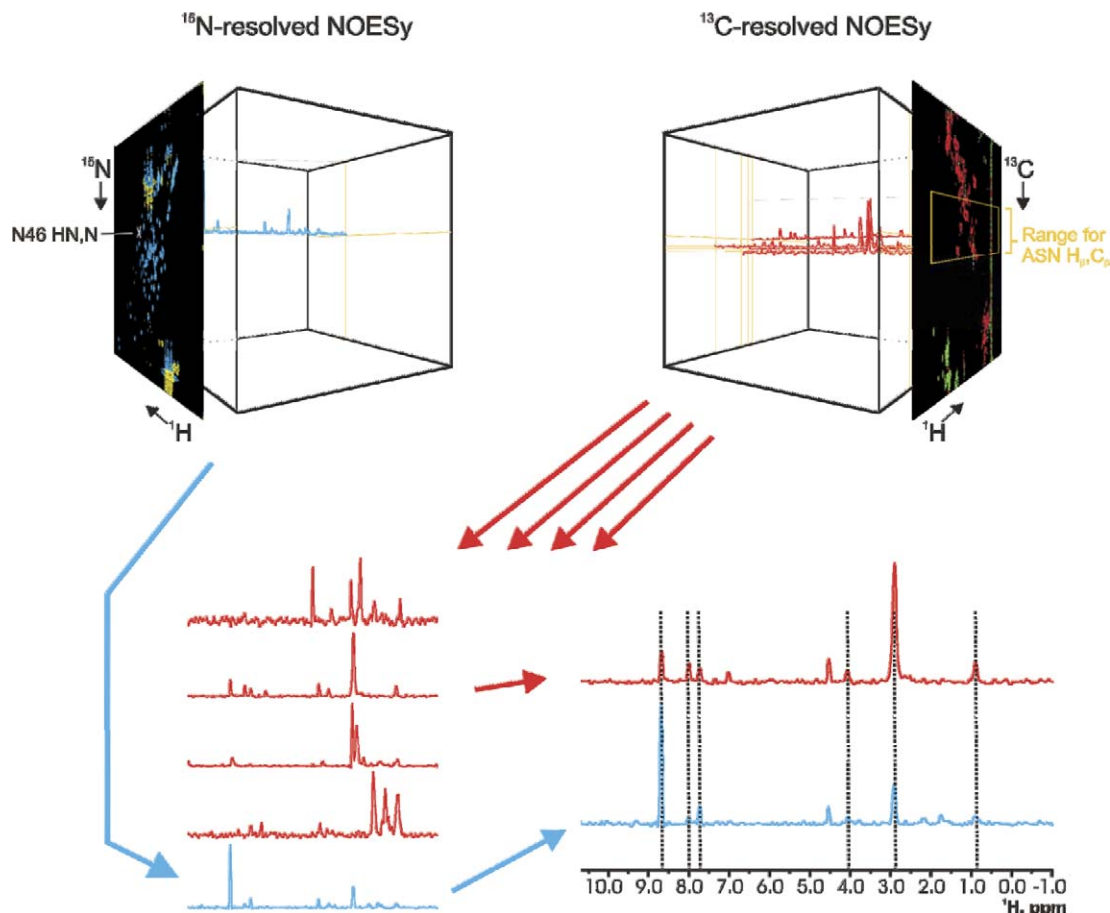


Fig. 1. Schematic showing the general approach to assigning side-chains using ^{15}N - and ^{13}C -resolved spectra. Initially spectral lines of interest are isolated from ^{15}N -resolved and ^{13}C -resolved 3D spectra (top). Subsequently the spectral lines are compared point-to-point (Eq. (1)) for similarity (bottom). Spectral lines that have coordinated amplitude distribution and appropriate $[\text{}^1\text{H}-\text{X}]$ COSY frequencies are then grouped into spin systems. In this example the spectra are a ^{15}N -resolved- and a ^{13}C -resolved NOESy of mMjCM. The amide spectral line (blue) was previously assigned to ASN 46 and the carbon-resolved spectral lines (red) are candidates for assignment to ASN 46 $\text{H}_\beta/\text{C}_\beta$ atom pairs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

of the ^{15}N -resolved NOESy, while simultaneously considering likely resonance frequencies based on empirically derived average chemical shifts as well as the few side-chain chemical shifts that were obtained with the backbone assignments. $[\text{}^1\text{H}, \text{}^{13}\text{C}]$ -COSY [64,65] Slices from a ^{13}C -resolved NOESy or HCCH-TOCSY that have relatively similar cross-peaks to a slice from the ^{15}N -resolved NOESy often signify that the $[\text{}^1\text{H}, \text{}^{15}\text{N}]$ -resolved resonances of the $\{\text{}^1\text{H}, \text{}^1\text{H}\}$ -NOESy slice and the $[\text{}^1\text{H}, \text{}^{13}\text{C}]$ -resolved resonances of the ^{13}C -resolved spectrum correspond to atoms that belong to the same residue of the protein under study. Since the $[\text{}^1\text{H}, \text{}^{15}\text{N}]$ -resolved resonances are already assigned to a specific residue during the backbone assignment process, the $[\text{}^1\text{H}, \text{}^{13}\text{C}]$ -resolved resonances can also be assigned, provided that they fit within the empirically known resonance frequency range of appropriate atoms in the specific residue's side-chain.

The major advantage of this approach is that it involves only a few additional spectra beyond what is needed for obtaining backbone resonance assignments, and these spectra are of relatively high sensitivity and resolution.

The major disadvantage to this approach is that no certain answers to the side-chain assignment problem can be obtained, as the NOESy data (for assignment purposes) only contains “soft” correlation information and therefore can be misleading. Thus, only relative certainty rather than absolute is possible, and confirmation of the results must await further analysis (i.e., structure calculation). Despite this limitation, it has proven reliable, and we argue, might be for that reason a desirable strategy for automated side-chain resonance assignment.

However, given the complexity and the “nebulous” nature of the side-chain data, it is not surprising that assigning side-chains effectively by software has proven elusive. To mimic the spectroscopist's approach, any such software must, in effect, be able to sort through a myriad of possible combinations of associatable data points, while relying solely on relative criteria for its decisions. Furthermore, since the side-chain detection spectra are generally less sensitive and resolved than the corresponding backbone spectra, the program must also be able to discern relative certainty considering all alternative assignment possibilities

in light of limitations in sensitivity and spectral overlap. As a final consideration, since the user cannot be wholly certain even of peak positions in the data prior (and sometimes even after) resonance assignment, it would be highly advantageous if the assignment software did not require any knowledge of peak positions in the side-chain assignment spectra (which would have to be provided by the user) prior to its operation.

To be able to fulfill these requirements, a slight paradigm shift as to how the side-chain assignment problem is viewed is useful. Specifically, a computer and a human necessarily must use different mechanisms to compare slices of 3D spectra. A human spectroscopist compares slices of 3D NMR spectra by placing them side-to-side and comparing peak positions. This is because a human visual cortex groups the data it views into useful elements (peaks) whose centers can be compared. Since computers do not inherently group spectra into peaks, instead different slices within 3D spectra must be viewed as spectral lines and compared with a series of mathematical and logical functions. The closer such a series of functions comes to reproducing the results of visual inspection by a spectroscopist, the closer the overall results of the automatic assignment process that uses them can be to that of the human spectroscopist. For a detailed description to the series of functions used by SideLink, see Section 2.6.

Other than this wetware-to-software methodological shift in comparison of spectral lines, the remainder of the spectroscopic analysis can be done by substituting computer-based simulated logic for human reasoning. The logic emulator embedded into SideLink is described below in Section 2.3 while its application to assigning side-chains is described in Sections 2.5 and 2.7.

2.2. CARA definitions

To describe SideLink, it is useful first to present a few key definitions inherited from SideLink's host program CARA ([66,67], www.nmr.ch). In CARA, a "spin" refers to a position in NMR spectra. It has a specific frequency and can be assigned to a specific atom of the molecule being studied. Cross-peaks in multidimensional NMR spectra, therefore, signify the interaction of two or more spins.

A set of such spins that are assignable to the same residue of the molecule under study is considered a "spin system." For most 2D and 3D NMR spectra, this means that the spins of a spin system are manifested by the frequencies of the various cross-peaks along a single line in the NMR spectrum as well as the frequency coordinates of the line in the other dimensions of the spectrum. In ^{15}N -resolved 3D spectra, such a line can be designated by a holding the amide ^{15}N and amide ^1H spin frequencies constant and varying the frequency in the remaining "cross-peak" dimension of the spectrum. Since such spectral lines may intersect inter-residue cross-peaks, a spin system may also contain spins that relate one spin system to spins of another spin system, such as seen for the $\text{C}_{\alpha-1}$ and $\text{C}_{\beta-1}$ cross-

peaks in 3D HNCACB experiments [68]. SideLink, in principal, does not need to make use of these inter-residue cross-peaks, so for the purpose of this discussion only, a spin system may be simply considered as all of the resonances of a single molecular residue.

2.3. Relative hypothesis prioritization (RHP) logic emulator

To describe how SideLink processes a side-chain assignment problem, it is useful to first describe at a more abstract level the logic emulator we have developed that is at the core of the program. While this engine is similar to the one used in the backbone assignment program AutoLink, it is considerably more complex, reflecting the increased complexity of the problem for which it was designed. Since SideLink, like AutoLink, uses relative hypothesis prioritization, the reader is directed toward the publication introducing the AutoLink program for an in-depth description of the basic RHP algorithm. Here there will be presented a review of the RHP process, including a few key definitions, which will highlight the additions made for SideLink.

The RHP process (Fig. 2) begins by first dividing the overall combinatorial problem into comparable subcomponents. For the purpose of this algorithm, a "hypothesis" refers to a single such subcomponent. Each hypothesis consists of one or more "criteria," which distinguish the hypothesis from other hypotheses, and a "priority score," which is a measure of the relative "likeliness" of that hypothesis to belong to the final solution. An example of a hypothesis that would be relevant to SideLink, for example, is the statement "There is a ρ relative probability that spin i belongs in spin system j " (designated as spin $i \rightarrow$ spin system j , ρ). Here "spin i " and "spin system j " would be considered the criteria of the hypothesis, and the priority score would be ρ . There is no requirement for hypotheses to have like criteria or even the same number of criteria.

Once generated, the hypotheses are grouped into sets (called "hypothesis sets"). These sets are the functional decision block for the RHP algorithm, as all component hypotheses must be either accepted as true or rejected as false together. Thus, the hypotheses within a single hypothesis set are evaluated in a manner consistent with a Boolean AND function [69]. All of the hypothesis sets within a combinatorial problem need not contain the same number of hypotheses, and some of the hypotheses of different sets may be identical. There are also no specific requirements that the hypotheses within a set must meet with regard to compatibility, as the RHP processing (described below) will allow them to be co-accepted even if they would appear to be contradictory. This is a particularly useful feature as it allows the RHP processor to work with "functional equivalence" (described further later). The various strategies for grouping hypotheses into sets allow every kind of inclusive hypothesis relationship to be encoded in an RHP-processable form. For example, if hypothesis A and hypothesis B are grouped into a set, but no other set exists

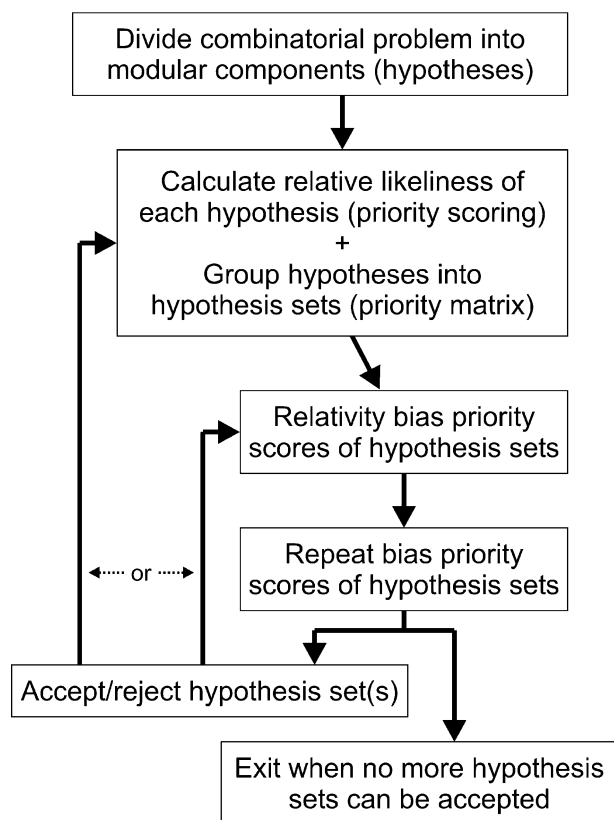


Fig. 2. Diagram of the general RHP process. The RHP process proceeds in cycles. During each cycle a decision is made to either accept or reject a hypothesis set based on its priority score. The initial priority scores for each hypothesis of each set are based on the relative likelihood of the hypothesis. These priority score can be calculated prior-to, concurrent with, or after grouping into hypothesis sets which are the fundamental decision units of the process. After the hypothesis scores and the hypothesis set groupings are determined, the priority score for each hypothesis set is defined as the score of the lowest scoring hypothesis component of the set. Decisions to accept or reject hypothesis sets are not made based on the initial priority scores. Rather the scores are first biased based on the uniqueness of the hypotheses in the sets (relativity biasing) and also based on the acceptance/rejection of prior hypothesis sets (repeat biasing). The process ends when no more hypothesis sets can be accepted or rejected either due to their inherent likelihood, or due to their lack of certainty caused by the existence of competing exclusive hypotheses. The process is described in more detail under Section 2.3.

that includes either A or B, then hypothesis A and B are mutually inclusive—i.e., both must be evaluated as true or neither can be. Alternatively, if another hypothesis set exists that contains hypothesis A but not hypothesis B then a one-way inclusive relationship is defined [70]. Hypothesis A can be accepted without B, but B cannot be accepted without A. Other kinds of inclusive hypothesis relationships can obviously be defined if more than two hypotheses are involved.

The use of hypothesis sets instead of single hypotheses is a significant advancement of SideLink over AutoLink. AutoLink, in comparison to SideLink, needs no inclusive hypothesis relationships, so its RHP process can be viewed as the same as that for SideLink except that all hypothesis sets would include exactly one hypothesis.

Once the hypothesis sets have been defined, they are then included into a “priority matrix” (see Fig. 3A). The priority matrix is the smallest functional data block that contains enough information to encode an entire combinatorial problem. Processing of the priority matrix requires that the compatibility of the various quantum decision blocks, the hypothesis sets, be assessed. Hypothesis sets may be considered either mutually compatible (Boolean “OR” relationship) or incompatible (Boolean “XOR” relationship) depending on their component hypotheses. In light of the hierarchical representation of the combinatorial problem in the priority matrix, it is useful to define a hierarchical set of compatibility rules to use in evaluating it. These rules are shown in Table 1.

The lowest level rules define how the criteria from different hypothesis are evaluated. For SideLink, criteria between hypotheses are considered to be in conflict if they are the same. Not all criteria need be considered as potentially conflicting—criteria can be defined as either exclusive or non-exclusive. The use of these rules will become clearer in the next paragraph where compatibility of hypothesis is considered.

The following compatibility rules between hypotheses only apply to hypotheses that are in different hypothesis sets. As previously mentioned, hypotheses within the same set are considered as defined to be compatible, no matter what their criteria are. For hypotheses in different sets, the hypotheses are considered to be incompatible if they contain one or more incompatible criteria—i.e., criteria that have the same value and are defined to be of the exclusive type. To exemplify this consider the previous hypothesis example “spin $i \rightarrow$ spin system j .” If the first criterion, “spin i ” is defined to be exclusive, then this hypothesis will conflict with any other hypothesis that groups spin i into a spin system other than j (except for those within the same hypothesis set, of course). On the other hand, if the second criterion is defined to be non-exclusive (as makes sense since many spins can belong to the same spin system), then the above hypothesis will be considered compatible with other hypotheses that group other spins into spin system j .

An additional rule for hypothesis comparison must be considered, that of redundancy. Hypothesis A is considered redundant with hypothesis B if all of the defined criteria of hypothesis A have identical values defined for hypothesis B. In the case of problems like SideLink where all of the hypotheses contain the same number of defined criteria, this means that this redundancy rule is commutative (if hypothesis A is redundant with hypothesis B, then hypothesis B is redundant with hypothesis A). For other kinds of problems, where the number of criteria defined for each hypothesis may vary, this commutative relationship is not necessarily true.

With these hypothesis compatibility rules defined, it is now possible to apply a single rule to evaluate the compatibility of hypothesis sets: hypothesis set A is considered compatible with hypothesis set B if every hypothesis in

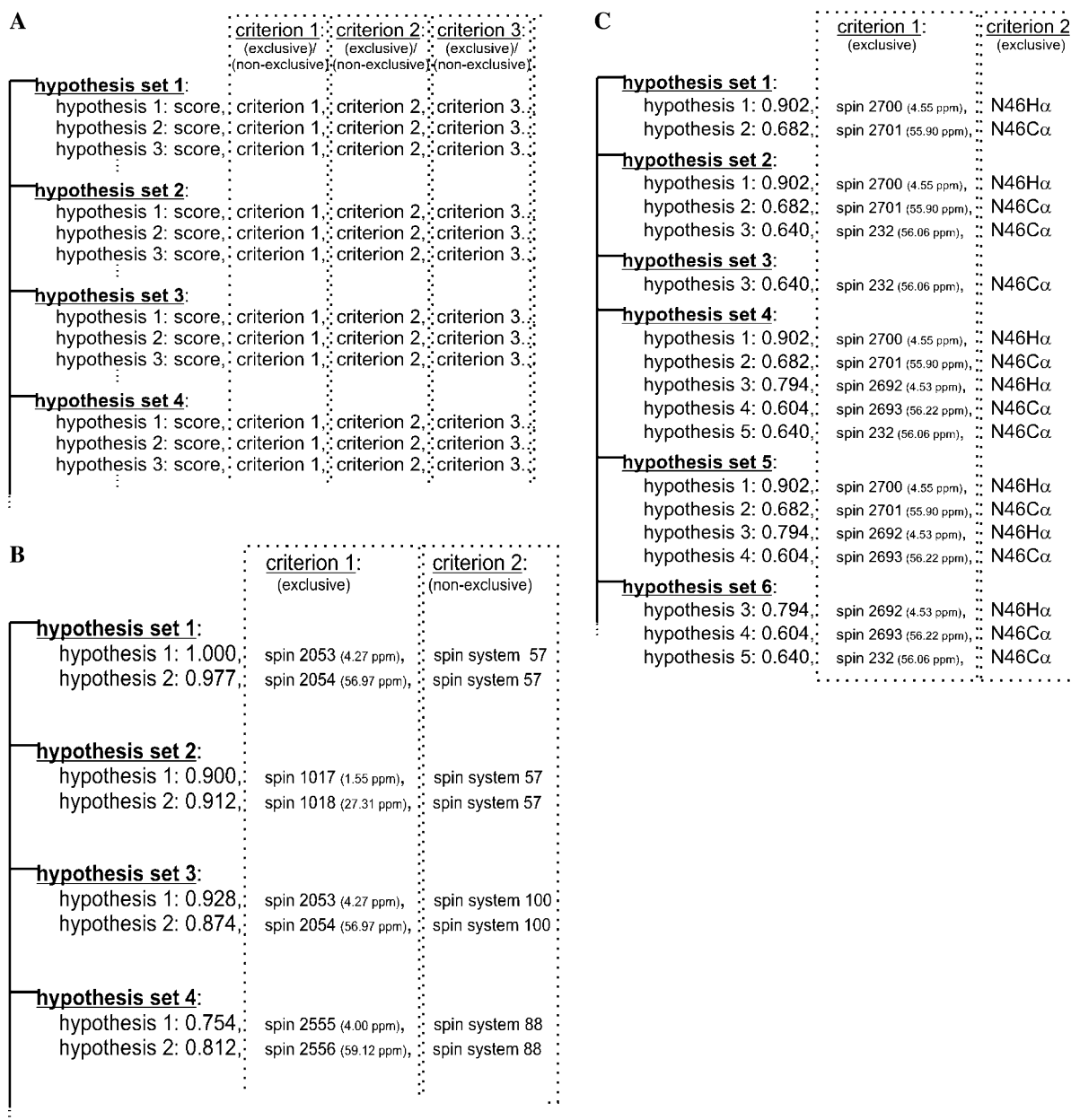


Fig. 3. (A) General model of a priority matrix. Note that a priority matrix can contain any number of hypothesis sets, each hypothesis set may contain any number of hypotheses, and each hypothesis can contain any number of criteria. Such a matrix can be viewed in terms of Boolean logic: The hypotheses of any given hypothesis set may be considered as a logical AND relationship since the entire group must be either accepted or rejected coordinately. Hypothesis sets with no incompatible criteria (see Table 1 and Section 2.3) are related by an OR function while those with incompatible criteria are related by an XOR function. Since all of the basic binary Boolean functions are represented, and since the hypothesis sets may contain overlapping hypotheses, virtually any complex logic relationship can be defined with an appropriate priority matrix. (B) Model of a priority matrix encoding spin → atom fitting within a spin system (“inner combinatorial problem”). In the interest of brevity only a subset of the hypotheses are shown. This example shows many of the basic tenants of the hypothesis compatibility rules. For example, there are no hypothesis sets that contain spin 2700 without also including 2701 (assigned to N46H_α and N46C_α, respectively). This imbues a mutual AND relationship between the two spins, since neither one can be accepted without also accepting the other. Spin 232, on the other hand, exists singly in a hypothesis set, and therefore can be accepted (matching N46C_α) without coordinate acceptance of another hypothesis. The relationship between hypothesis set 3, which assigns only spin 232 to N46C_α, and hypothesis set 1 which assigns spin 2701 to the same atom is exclusive (XOR), since acceptance of either set precludes acceptance of the other set. If spin 232 is close to spin 2700 in chemical shift, however, another hypothesis set can be created, i.e., hypothesis set 2, which allows coordinated assignment of both spins to the same atom. (Note that acceptance of hypothesis set 1 or 3 does not preclude the acceptance of set 2, since all coincidental criteria are part of identical hypotheses.) A logical OR relationship would be exemplified by the inclusion of other hypothesis sets that relate other spins to other atoms (not shown). Since the spins are different and the atoms are different, the hypotheses sets can be accepted or rejected independently of each other without exclusion. (C) Model of a priority matrix encoding spin → spin system grouping (“outer combinatorial problem”). Here, we see the use of a non-exclusive criterion, spin system. Since multiple spins can fit into a spin system, the acceptance of one spin into a spin system doesn’t necessarily preclude the addition of another spin into the system. There is an additional consideration here, however, that all of the spins added to a spin system must be able to simultaneously fit into the system. This condition is not directly handled by the encoding of the priority matrix criteria, but rather is implemented in the RHP processing by re-evaluating the priority scores of each hypothesis set acceptance/rejection round based on the fit of the spin systems to their assigned residue of the protein sequence.

Table 1
Hierarchical compatibility rules for priority matrix evaluation

Lowest level

Rules for criterion compatibility

- (1) Criteria can be considered exclusive or non-exclusive.
- (2) An exclusive criterion from hypothesis A is considered to be incompatible with the equivalent criterion of hypothesis B if the criteria are identical.

Intermediate level

Rules for hypothesis redundancy

- (1) Hypothesis A can only be redundant with hypothesis B if the number of criteria in hypothesis A must be less than or equal to than the number of criteria in hypothesis B.
- (2) Hypothesis A is redundant to hypothesis B if every criterion of hypothesis A is incompatible with the equivalent criterion of hypothesis B.

Rule for hypothesis compatibility

- (1) Hypothesis A is incompatible with hypothesis B if some but not all of its exclusive criteria are incompatible.

Highest level

Rule for hypothesis set compatibility

- (1) Hypothesis set A is incompatible with hypothesis set B if any of the hypothesis in A conflict with any of the hypothesis in B, and the conflicting criteria of the conflicting hypotheses are not present in any hypothesis of hypothesis set A that is redundant with a hypothesis of hypothesis set B or vice versa.
-

set A either does not conflict with the hypotheses in set B, is redundant with a hypothesis in hypothesis set B, or if any conflicting criterion value of the hypotheses within set A also occurs in another hypothesis in set A that is redundant with a hypothesis in set B. An example will clarify this definition. If hypothesis set A contains only a signal hypothesis, “spin 1 → residue 5 H_α,” and hypothesis set B contains only “spin 2 → residue 5 H_α,” then set A would be incompatible with set B as long as the first criteria was defined as exclusive. This makes sense in this example because the spins, which are associated with different frequencies, cannot both be assigned to the same atom, since each probably should have only one frequency. If, however, hypothesis set A also contained “spin 2 → residue 5 H_α” (the same hypothesis as in set B and therefore a redundant hypothesis), then the two hypothesis sets would be considered compatible. In this example, spin 1 and spin 2 are given the possibility of being “functionally equivalent” assignments to residue 5 H_α, possibly because their frequencies are very close. It should be noted that, although inclusion of such hypothesis sets allows the assignment of functional equivalence, it does not automatically guarantee it. In the above example, it is still possible that only spin 1 or spin 2 or even neither of the two be assigned to residue 5 H_α, provided other hypothesis sets exist in the same priority matrix that allow separate or alternative assignments.

Given that the structure of the priority matrix and its associated compatibility rules can represent all Boolean operations, it is possible to encode virtually any type of combinatorial problem into the matrix in a manner that can be evaluated by RHP-logic (see the next section). However, one should be cautioned in the case of problems that involve significant symmetry in the possible solution set, as encoding them into the priority matrix can involve creating a number of hypotheses that is related to the factorial of the size of the problem. In such cases, a scoring function that biases in favor of one of the symmetry-related solutions is advisable to ensure that one of the equivalent solutions is selected from among the possible permutations.

A major advantage of RHP for the evaluation of combinatorial problems is a drastic reduction in the number of permutations that must be considered compared to the actual number of permutations within the solution space of any given problem, even if the problem is highly non-monotonic. For example, consider a combinatorial problem with only 10 elements which must be ordered. For a problem of even this small size there are 9! (362,880) permutations. RHP analysis, however, requires the construction of a priority matrix with at most 90 hypotheses, making processing of the problem considerably more tractable. NMR problems often have on the order of 100–200! permutations in the solution space. This number may be a little misleading, however, because most of these permutations can be quickly eliminated as obviously incorrect. Despite this, most NMR assignment problems will still have >1,000,000 possible solutions that must be considered. RHP allows such large problems to be handled relatively quickly as the process time and memory usage is more proportional to the abstract concept of the “complexity” of the problem rather than the size.

2.4. Evaluation of the priority matrix by relative hypothesis prioritization

RHP evaluation of a priority matrix is a cyclic process with each cycle divisible into an initial scoring phase, a biasing phase, and a hypothesis acceptance or rejection phase. In the initial scoring phase, each hypothesis in the priority matrix is assigned a priority score which reflects the likelihood that the hypothesis is true. This score is generally a function of the values of the criteria that define the hypothesis. In the case of more monotonic logic problems, this phase must only be executed once at the beginning of the first cycle, as the initial scores do not change from one cycle to the next. For highly non-monotonic logic such as that use for both backbone and side-chain resonance assignment problems, however, the priority score of each hypothesis depends not only on the criteria of the

hypothesis itself, but also on the acceptance or rejection of other hypotheses in the previous cycles. Thus, the priority scores must be re-calculated at the beginning of each cycle.

Once the individual hypothesis scores are known, the overall score for each hypothesis set is defined to be the score of the lowest-scoring hypothesis of the set, increased by an “insignificant factor” times the number of hypotheses in the set. The insignificant factor is a very small percentage used to ensure that among hypotheses sets that contain redundant hypotheses (but have the same overall score for the set), those with more hypotheses are slightly favored above those with fewer hypotheses and is necessary to insure convergence to a solution if any of the hypothesis sets contain functionally equivalent hypotheses.

These initial scores are not the final rating for each hypothesis set that is used to evaluate the priority matrix. Instead, the scores are adjusted according to “biasing potentials.” Though other types of biases are definable, both AutoLink and SideLink rely heavily on two, relativity bias and repeat bias, to evaluate their combinatorial problems.

2.4.1. Relativity biasing

Relativity biasing is a mechanism for weighting the hypothesis scores according to the uniqueness of the criteria of the hypothesis within each hypothesis set. For each criteria of a hypothesis set, the priority matrix is scanned for the next better scoring incompatible hypothesis set that conflicts at that criterion and the next worse incompatible hypothesis set, also conflicting at that criterion. One relative score is calculated for each criteria of the hypothesis set, and the best one is used as the relativity biased score for the set. To calculate each criterion’s relative score, the hypothesis score is first multiplied by the difference between its score and the score of next worse incompatible hypothesis set. However, if this relative score exceeds the relative score of the next better hypothesis set that conflicts with the criterion, then the score is set to the relative score of the next best hypothesis set minus the insignificant factor. This new score is then incremented by the insignificant factor squared times the number of hypothesis in the set to maintain the slight bias in favor of hypothesis sets involving more hypotheses over those with fewer.

2.4.2. Repeat biasing

Repeat biasing accomplishes two functions. First, it prevents RHP processing from ever getting caught in unending loops. Second, it gives the RHP evaluation the ability to assess the problem in terms of relative certainty, and reject any component of a solution that cannot be known to a specific degree of certainty. For repeat biasing, each hypothesis set priority score is reduced depending on how many times that hypothesis (or another hypothesis set involving the same criteria as the first set) has been either accepted (positive repeat biasing) or rejected (negative repeat biasing). Since SideLink, like AutoLink, uses only positive repeat biasing, this discussion will be limited to that form. For positive repeat biasing the overall priority

score for each hypothesis set is repeat biased once for each criterion in the set. The repeat bias for any given criterion is governed by:

$$score' = score \times rb^n, \quad (1)$$

where *score* is the priority score before repeat biasing, *score'* is the priority score after repeat biasing, *rb* is the user-defined repeat bias (ranging from 0 to 1), and *n* is the number of times the hypothesis has been previously accepted. As a special case, repeat biasing based on non-exclusive criteria only affects other criteria that have the same exclusive criteria as well as the same non-exclusive criterion. Since a hypothesis set can contain several values for any given criterion, the hypothesis set score is only modified for each criterion using the criterion value that causes the strongest bias (i.e., causing the largest reduction of the priority score). The overall effect of this bias is that, upon acceptance of a hypothesis set, the priority score of that set is reduced in subsequent RHP cycles. This may allow an incompatible hypothesis that was previously lower in score to increase in relative priority in subsequent rounds. In fact, any two competing hypotheses sets whose scores differ by a smaller fraction than the repeat bias control parameter will alternately be accepted and rejected until either one becomes impossible due to the acceptance of another hypothesis set or until the scores of both hypothesis sets are reduced below a critical limit (also user defined) which will cause both hypothesis sets to be rejected (considered “irresolvable”). In effect repeat biasing causes the hypothesis sets that are close in relative priority score to be considered in the context of various combinations of other hypothesis sets to determine if a consistent set of sets can be discerned.

2.4.3. Hypothesis acceptance/rejection

The main decision-making stage of the RHP cycles is the hypothesis acceptance/rejection phase of the cycles. After the initial scores are modified by the biasing potentials, they are used as a measure of the “acceptability” of a hypothesis set. Each RHP cycle can be defined as positive or negative, depending on whether hypotheses are to be accepted or rejected in that cycle. It should be noted that in positive cycles, it is possible that a hypothesis set be rejected if it is incompatible with a better-scoring hypothesis set. In each acceptance/rejection cycle, the priority matrix is scanned in order of hypothesis set priority in order. In positive cycles, the highest scoring, not-previously-accepted hypotheses are identified. A user-defined number of these hypotheses is then accepted, with one additional consideration: no hypothesis can be accepted if it is incompatible with a higher-scoring hypothesis.

For the negative cycles, the lowest-scoring, previously accepted hypothesis are identified and subsequently rejected. Obviously, in order for the evaluation to proceed toward a solution, the overall number of hypotheses accepted must exceed the number of hypotheses rejected. Negative cycles are generally only useful for processing

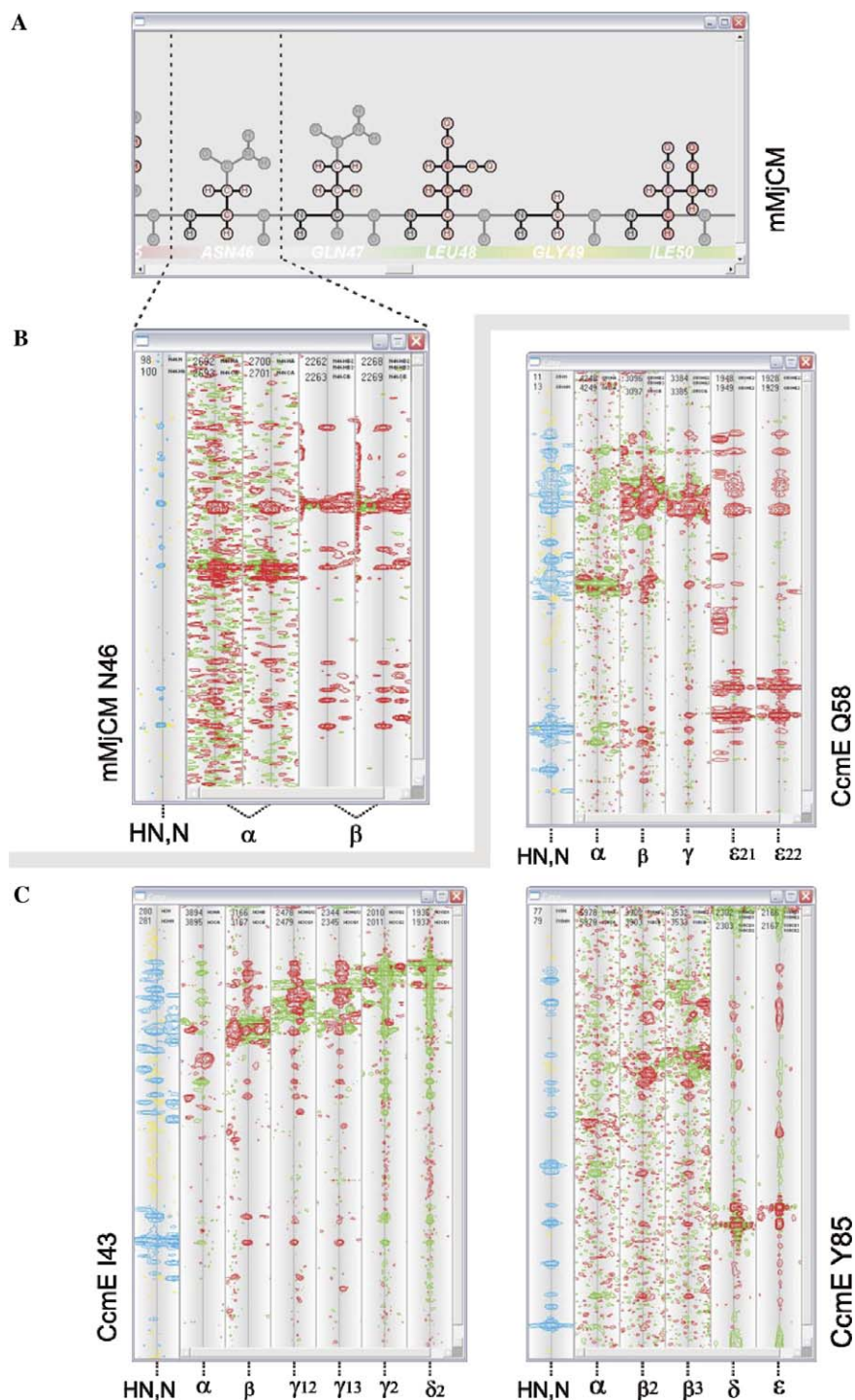


Fig. 4. (A) Demonstration of residues displayed in SideLink user interface. For the purpose of SideLink, each residue represents one spin system. White-to-red color coding of each residue's atoms allows the user to quickly navigate through the results. For each atom, both the chemical shift fit (indicated by the color near the center of atom) and comparison score (indicated by the color near the edge of the atom) are accessible. Further information, including a display of the spectral lines grouped into the spin system, and a display of all possible spectral lines assignable to each atom, is readily available through right-click-driven functions. (B) Spectral lines from a ^{13}C -resolved NOESY representing spins grouped into spin system 46 (mmJCM Asn 46). Grouping of such spectral lines together by RHP logic into a spin system (the outer combinatorial problem—see Section 2.5) is the overall goal of the SideLink program. Note that low signal-to-noise ratio (i.e., for spins 2766 and 2767) does not prevent assignment of the spectral line to the correct spin system and that redundant spectral lines are assigned to the same atoms. Atom assignments are in fact also deduced by RHP logic (inner combinatorial problem—see “Encoding the side-chain assignment problem into the priority matrix” subsection “Spin group \rightarrow residue fitting”). An example of the priority matrix created to assign these spins (C_α components shown only) is shown in Fig. 2B. (C) Spectral lines groupings for I43, Q58, and Y85 of CcmE. For I43 and Q58 the spectra lines (including the side-chain amino resonances of Q58) were obtained from a single 3D NOESY containing both ^{13}C - and ^{15}N -resolved NOEs. For Y85 the spectral lines assigned to δ and ϵ positions were obtained from an additional ^{13}C -resolved NOESY optimized for aromatic resonances. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

highly non-monotonic problems where the initial priority scores at the start of the cycles are highly influenced by the hypotheses accepted and rejected in previous cycles.

Processing of the priority matrix is completed when there are no more not-accepted hypotheses in the priority matrix which are not in conflict with better-scoring accepted hypotheses and with biased scores above the user-defined cutoff. This cutoff is generally set to a very low value that is consistent only with an extremely improbable hypothesis set.

There are two ways of interpreting the final outcome of the RHP processing depending on whether one is interested in finding out whether a reasonable solution exists that can account for all criterion values in a combinatorial problem, or whether one is only interested in relatively certain conclusions. In the latter case, the list of accepted hypotheses contains the final solution, as all undeterminable hypothesis sets have been excluded by being repeat biased below the critical threshold. In the former case, however, all hypothesis sets that are below threshold due only to repeat biasing and the presence of an alternative reasonable hypothesis set are re-examined and the best scoring of these is accepted. The resulting hypothesis set can then be regarded as one possible reasonable solution. Though AutoLink was only interested in determining relatively certain solutions, both of these interpretation methods are used by SideLink to solve side-chain assignment problems (as will be described subsequently).

2.5. Encoding the side-chain assignment problem into the priority matrix

To encode the side-chain assignment problem into the priority matrix, it is necessary to format it into one or more combinatorial logic problems. In our implementation, the problem has been divided into an outer combinatorial problem and an inner combinatorial problem, with one inner combinatorial problem used to calculate the priority score for each hypothesis in the outer combinatorial problem.

The outer combinatorial problem can be viewed as figuring out which spectral lines of a spectrum belong in the same spin system (Fig. 4). Since the program has been

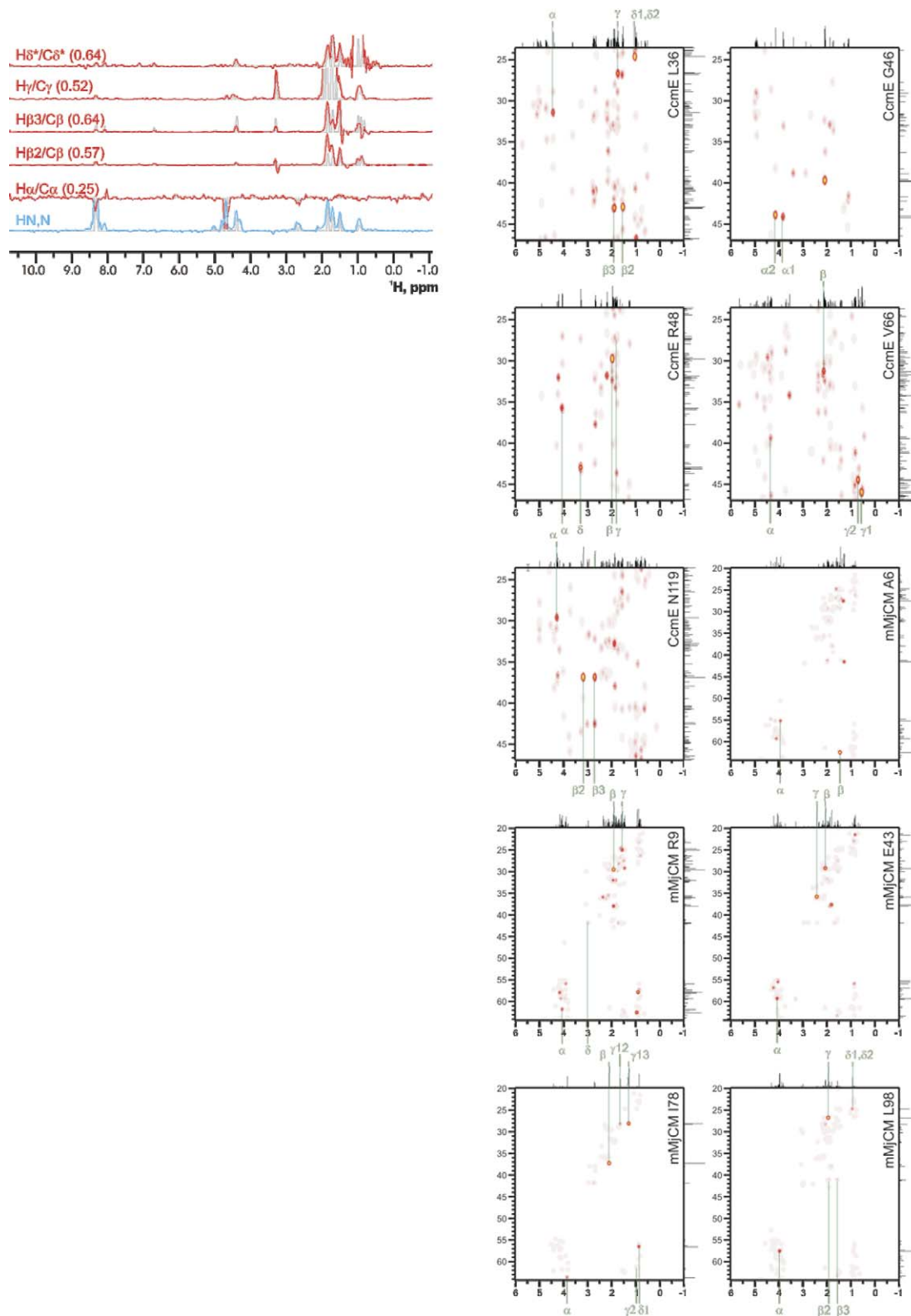
developed using 3D data, this discussion will focus on this particular case for clarity. It is relatively straightforward to generalize the application of the algorithm to spectra of different dimensionality. For clarity, a spectral line, in this case, refers to a specific line of points in a 3D spectrum which are indexed by two spins (frequencies) in the [^1H -X] COSY (where X = ^{15}N , ^{13}C) dimensions of the spectrum. While in the NOE dimension of the spectra the peaks may be from atoms of separate residues, the [^1H -X] spins designating the line always belong to the same spin system (for proteins), and this spin system is the one with which SideLink seeks to group the spectral line. Since the backbone resonances are already assigned prior to running the program, each spin system (at least the ones that SideLink is interested in) already has at least one spectral line associated with it, that of the amide ^1H and amide ^{15}N of the ^{15}N -NOESY. Comparison of other spectral lines to the amide line is one of two factors used in determining the priority scores of the outer combinatorial problem. The other factor is the hypothetical “fitness” of the [^1H , ^{13}C]-correlation-like chemical shifts of the spectral lines with regard to the expected chemical shifts of the spin system since its residue type is known (fitting spins of unknown identity into specific atoms of the spin system is the source of the inner combinatorial problem and will be discussed in detail later). This fitness score takes into account the fit of the previously grouped spins of the spin system as well as the new spins hypothetically to be added to the spin system. Thus, the hypotheses of the outer combinatorial problem are all of the form “spin \rightarrow spin system” with priority scores factoring in the comparison of the spectral lines in the NOE dimension and the chemical shift values of the spectral line in the [^1H -X]-correlation dimensions, where X is ^{15}N or ^{13}C . Of course, for 3D NOESY spectra, each hypothesis set will contain two such hypotheses, since there are two frequencies designating each spectral line. See Fig. 3B for an example of a priority matrix for the outer combinatorial problem. In these outer combinatorial problem hypotheses, the first criterion, “spin,” is treated as exclusive, while the second criterion, “spin system” is non-exclusive. This is because each spin can only belong to one spin system while several different spins can belong to the same spin system.

Fig. 5. (Top) 1D spectral slices assigned to CcmE L36. The unprocessed amide spectral line is shown in blue while the unprocessed [^1H , ^{13}C] spectral lines are shown in red. The relative amplitudes are shown as they are in the spectra, with the highest amplitude peaks cut off to make the low amplitude peaks visible. The processed spectral lines are displayed in grey beneath the unprocessed spectral lines, with their amplitude scale = 1/20 of the amplitude of the unprocessed spectral line. Listed above each spectral line is the assignment within the L36 spin system and the relative correlation score of the spectral line to the amide line. (Right) Comparison plots of various amide spectral lines to [^1H , ^{13}C]-resolved NOESYS optimized for aliphatic resonances. For each plot, a box is shown corresponding to the [^1H , ^{13}C] projection of the spectrum (^1H resonances are shown on the horizontal axis and ^{13}C frequency on the vertical axis, both in units of ppm). Points within the box are color coded with comparison scores >0 fading from red to darker red to gold with increasing relative comparison scores. For each point of the plot a black line perpendicular to the edge is drawn with the length corresponding to the highest comparison score at the corresponding frequency. Indicated in green are the positions of the [^1H , ^{13}C]_{*i*} assignments for the spin system (where *i* = $\alpha, \beta, \gamma, \delta, \epsilon, \dots$). Note that while the resonances of the intra-residue spins correspond to spectral lines with relatively high comparison scores, for each spin system there are also several other spectral lines which are not part of the spin system that have significant comparison scores. This is as expected for the NOESY spectra due to the presence of inter-residue cross-peaks. These spectral lines are generally not assigned to the wrong spin system by SideLink because their [^1H , ^{13}C] frequencies are not compatible with the given residue and/or their comparison with their correct spin system is higher than the comparison scores to the wrong system. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

2.6. Spectral line comparison

The comparison of spectral lines is a function that is based on the point-wise multiplication of the two spectral lines for all points that are in their overlapping ppm range.

$$C(\{1\}, \{2\}) = \left[\frac{\sum_{x=1}^n \left[\frac{\{1\}_x}{I_{(1)x}} \times \frac{\{2\}_x}{I_{(2)x}} \right]}{n} \times (\text{cdf}_{1 \rightarrow 2} \times \text{cdf}_{2 \rightarrow 1})^{\frac{1}{2}} \right]^{\frac{1}{2}}, \quad (2)$$



where $\{1\}$ and $\{2\}$ are the spectral lines to be compared, and n is the number of points in common. $\text{cdf}_{1 \rightarrow 2}$ is a “cross diagonal factor” and has the value of 0.5 if spectral line $\{2\}$ has an amplitude greater than the user-defined threshold at the chemical shift of the diagonal resonance of spectral line $\{1\}$ and 0 if it does not. Likewise, $\text{cdf}_{2 \rightarrow 1}$ is a comparable factor related to the amplitude of spectral line $\{1\}$ at the diagonal chemical shift of spectral line $\{2\}$.

The $I_{\{?\}x}$ terms of this equation refer to the average amplitude of all of the spectral lines for the point x in the cross-peak dimension of the relevant spectrum. Division by these terms is analogous to AutoLink’s score density compensation. In effect, it causes parts of the spectral lines that have high density in many spectral lines to be reduced in importance when compared parts with high density in relatively few spectral lines. Fig. 5 demonstrates the comparison of various spectral lines from a $[^1\text{H}, ^{13}\text{C}]$ -correlated NOESY (optimized for aliphatic resonances) with amide spectral lines from a $[^1\text{H}-^{15}\text{N}]$ -correlated NOESY.

Prior to the point-wise scoring, the user can optionally process the spectral lines with filters, which include “peak compression,” “amplitude threshold,” and “local scaling.” These filters may be used in any combination. For a demonstration of comparable spectral lines and pre-processing filters see Fig. 6A.

2.6.1. Peak compression

Uncompressed peaks which are at slightly different chemical shifts, but whose shoulders overlap will give a positive score >0 in the above equation, even if they are (to a human spectroscopist) obviously unrelated. To address this problem, the user can use peak compression to narrow the peaks in a spectral line to reduce shoulder overlap. What this means in terms of the program is that the spectral lines are deconstructed based on the average peak shape for the spectrum prior to point-wise multiplication (see Fig. 5). The deconstruction used by SideLink is

based on maximum parsimony, fitting until all density of the spectral line is accounted for within the limit of the amplitude threshold (see next section). Though this may seem at first glance to be “peak-picking,” in effect it is not since the small deviation of each peak from the average peak shape causes the deconstruction algorithm to produce a distribution of peak amplitudes rather than a single point identification (see Figs. 6B and 7A).

2.6.2. Amplitude threshold

Each point of each spectral line is compared to a threshold and if the amplitude is lower than the threshold, then the amplitude is rounded to zero. This threshold may be either user-defined or dynamically calculated based on the local noise level of the spectral line. Amplitude thresholding primarily functions to reduce the effect of noise on the comparison scores. It is important to note that amplitude thresholding is most effectively used on spectra with a flat baseline.

2.6.3. Local scaling

Local scaling refers to the regional adjustment of the amplitude of spectral lines in the NOE dimension. Its primary function is to reduce the effect of regional amplitude disparities on the outcome of the spectral line comparison function. There are four options for local scaling, including “diagonal-two-zone,” “diagonal-2.5-zone,” “zero-to-zero,” and “single-zone” (see Fig. 7B). In each scaling mode the final scaled amplitudes of the spectral lines may be independent of other spectral lines (“non-global” local scaling), or alternatively, the scaling can maintain the overall relative amplitudes of the different spectral lines (“global” local scaling).

As the name implies, for diagonal-two-zone scaling, each spectral line is divided into two zones, one containing the diagonal peak and a second zone that contains the rest of the spectral line. The boundaries of the diagonal zone are defined by starting at the diagonal point and fol-

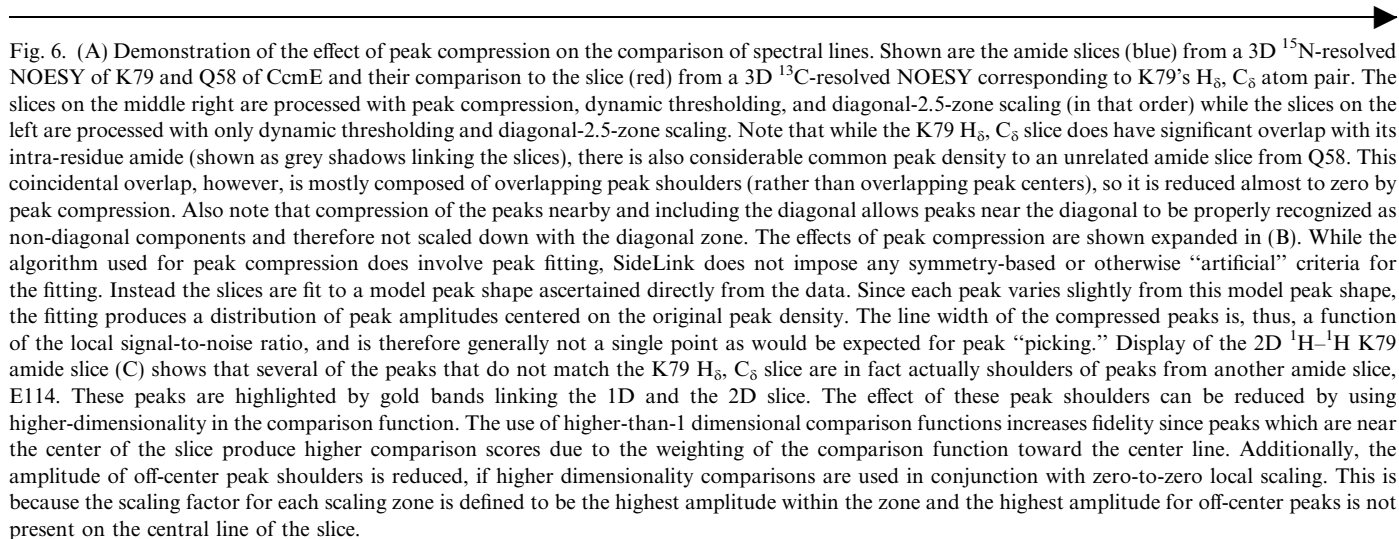
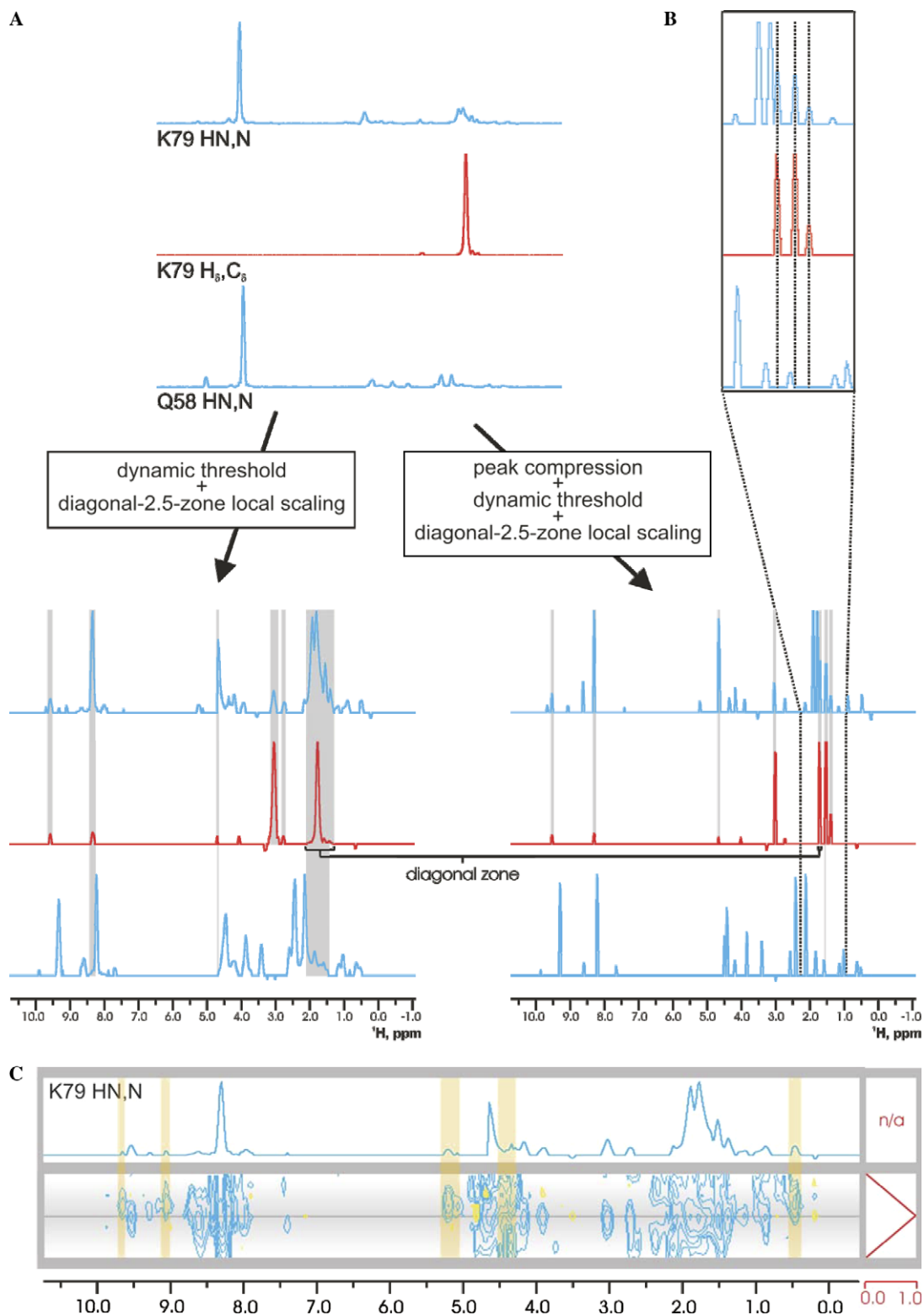


Fig. 6. (A) Demonstration of the effect of peak compression on the comparison of spectral lines. Shown are the amide slices (blue) from a 3D ^{15}N -resolved NOESY of K79 and Q58 of CcmE and their comparison to the slice (red) from a 3D ^{13}C -resolved NOESY corresponding to K79’s H_δ , C_δ atom pair. The slices on the middle right are processed with peak compression, dynamic thresholding, and diagonal-2.5-zone scaling (in that order) while the slices on the left are processed with only dynamic thresholding and diagonal-2.5-zone scaling. Note that while the K79 H_δ , C_δ slice does have significant overlap with its intra-residue amide (shown as grey shadows linking the slices), there is also considerable common peak density to an unrelated amide slice from Q58. This coincidental overlap, however, is mostly composed of overlapping peak shoulders (rather than overlapping peak centers), so it is reduced almost to zero by peak compression. Also note that compression of the peaks nearby and including the diagonal allows peaks near the diagonal to be properly recognized as non-diagonal components and therefore not scaled down with the diagonal zone. The effects of peak compression are shown expanded in (B). While the algorithm used for peak compression does involve peak fitting, SideLink does not impose any symmetry-based or otherwise “artificial” criteria for the fitting. Instead the slices are fit to a model peak shape ascertained directly from the data. Since each peak varies slightly from this model peak shape, the fitting produces a distribution of peak amplitudes centered on the original peak density. The line width of the compressed peaks is, thus, a function of the local signal-to-noise ratio, and is therefore generally not a single point as would be expected for peak “picking.” Display of the 2D $^1\text{H}-^1\text{H}$ K79 amide slice (C) shows that several of the peaks that do not match the K79 H_δ , C_δ slice are in fact actually shoulders of peaks from another amide slice, E114. These peaks are highlighted by gold bands linking the 1D and the 2D slice. The effect of these peak shoulders can be reduced by using higher-dimensionality in the comparison function. The use of higher-than-1 dimensional comparison functions increases fidelity since peaks which are near the center of the slice produce higher comparison scores due to the weighting of the comparison function toward the center line. Additionally, the amplitude of off-center peak shoulders is reduced, if higher dimensionality comparisons are used in conjunction with zero-to-zero local scaling. This is because the scaling factor for each scaling zone is defined to be the highest amplitude within the zone and the highest amplitude for off-center peaks is not present on the central line of the slice.

lowing the spectral line in either direction until the amplitude crosses zero, and then just before the amplitude crosses zero again. Once the diagonal zone has been defined, it is then linearly scaled such that its maximum amplitude is equal to either the maximum amplitude of the off-diagonal zone (global scaling) or 1 (non-global

scaling). This form of local scaling is particularly useful in limiting the effect of the diagonal on comparison scores, since the diagonal peak is often considerably larger than cross-peaks, while retaining the relative amplitudes of off-diagonal peaks, which are a source of valuable information in NOESY spectra.



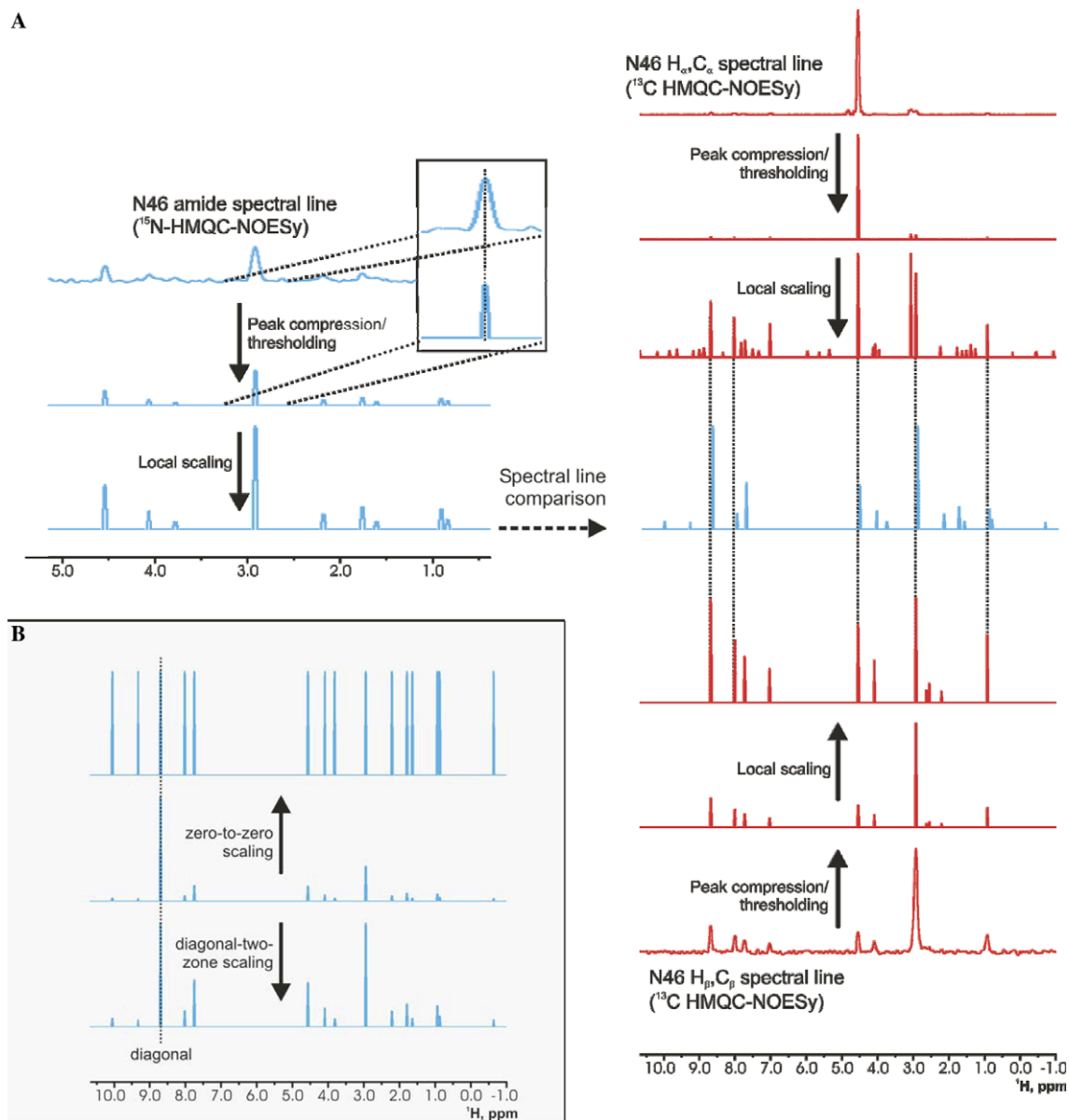


Fig. 7. (A) Demonstration of spectral line processing during spectral line comparison. Shown are spectral lines for mMjCM amide, $\text{H}_\alpha/\text{C}_\alpha$, and $\text{H}_\beta/\text{C}_\beta$ spectral lines. For each spectral line the unprocessed spectral line, the spectral line after peak compression (and non-dynamic thresholding), and the spectral line after peak compression and local scaling are shown. Alignment of the processed spectral lines demonstrates the presence of several relatively strong peaks in common across the three spectral lines. (B) Pictorial comparison of the effects of diagonal-two-zone and zero-to-zero local scaling for the mMjCM N46 amide spectral line from a ^{15}N -resolved NOESy. Note that with diagonal two-zone scaling (bottom) all points outside the diagonal zone maintain their relative amplitudes, while for zero-to-zero scaling (top) no such ratios are conserved. Diagonal-2.5-zone scaling (not shown) has a similar effect to diagonal-two-zone scaling except that an optional third scaling zone may be created if such a zone can be found where the highest point is comparable in amplitude to the diagonal intensity and significantly higher amplitude than any point outside the third zone or the diagonal zone. The additional optional zone for diagonal-2.5-zone prevents a single strong peak in the spectral line (for example the cross-peak relating $\text{H}_{\beta 1}$ to $\text{H}_{\beta 2}$ in the same side-chain) from outweighing all of the other non-diagonal peaks in the spectral line.

Diagonal-2.5-zone scaling is similar to diagonal-two-zone scaling except that an optional third scaling zone may also be defined if such a zone can be defined so that the amplitude of the zone is at least half the amplitude of the diagonal zone and at least twice the amplitude of any point outside of the new third zone or the diagonal zone. The primary function of defining this third zone is to prevent a single strong cross-peak, such as the $\text{H}_{\beta 1}$ – $\text{H}_{\beta 2}$

cross-peak (in a spin system where those resonances are not overlapped) from having an overly dominant effect on the spectral line's correlation scores.

Zero-to-zero scaling is similar to diagonal-two-zone scaling, except that several zones are defined, each zone boundary being set to points of the spectral line where the spectral line crosses zero. Each zone is linearly scaled to the same amplitude as the highest off-diagonal

zone (global scaling) or optionally to 1 (non-global scaling). This type of local scaling is particularly relevant to TOCSY-type spectra, where the relative peak amplitudes do not encode much information about peak identities.

Single-zone scaling is simply scaling the maximum amplitudes of each spectral line to 1 without any subdivision of the line. Obviously this is only relevant if global scaling is activated, as such an operation on isolated spectral lines out of the context of other spectral lines has no net effect.

2.6.4. Multidimensional comparison function

Rather than comparing one-dimensional spectral lines, SideLink can alternatively compare spectral regions of higher dimensionality (see Fig. 6C). Though the cross-peaks of interest to the program are generally arranged along a line in the NMR spectra, noise in the NMR spectra can often distort the amplitude of any given peak along a particular spectral line. To reduce the effect of noise on spectral line comparison, therefore, the point-wise multiplication of the spectral lines can be broadened to include the multiplication of points adjacent to the spectral line. However, this may inadvertently, especially in crowded parts of the spectra, cause bias to the comparison scores due to the presence of nearby adjacent spin systems in the NMR data. To reduce the effect of adjacent spin systems on spectral line scores the point-wise multiplication is, thus, modified by applying a linear penalty function to the spectral regions prior to point-wise multiplication, with amplitudes of the points that are further from the central line being proportionately reduced compared to those that are nearer the central line.

After all of the spectral line comparison scores have been calculated, they are subsequently linearly scaled such that the highest comparison score is always 1 (assuming that at least one non-zero score exists).

It is important to note that the approach to spectral line comparison method currently used by SideLink is not directly integral to the RHP logic used to solve side-chain assignment problems, and thus can be improved or replaced as newer methods are developed. Spectral line comparison scoring is, in fact, the primary limiting factor in the program's accuracy, largely due to artifacts present in the data.

2.7. Spin group \rightarrow residue fitting

Spin group \rightarrow residue fitting must be tested once per hypothesis per RHP round of the outer combinatorial problem. This fitting is the inner combinatorial problem mentioned under Section 2.5. Though this inner combinatorial problem is smaller than the outer problem, its encoding is somewhat more complex due to the need for functionally equivalent hypotheses.

Each hypothesis of the inner combinatorial problem is of the form “spin \rightarrow atom,” which obviously implies that

the program is trying to figure out which spin belongs to which atom of the spin system. The score of each hypothesis is calculated as:

$$1 - \frac{|\text{ppm}_{\text{avg}} - \text{ppm}_{\text{observed}}|}{\text{sd} \times \text{sdf}}, \quad (3)$$

where $\text{ppm}_{\text{avg}} - \text{ppm}_{\text{observed}}$ is the difference between the spin's chemical shift and the expected chemical shift, and sd is the standard deviation associated with the expected chemical shift. To save computation time, any hypothesis whose score is below a user-defined threshold (or equal to 0 if the threshold is set to 0) is not further considered.

Many of the spins in any group, especially those that are linked through heteronuclear magnetization transfers, must only be allowed to fit coordinately to atoms linked by a covalent bond. These are encoded in the priority matrix by including them exclusively in hypothesis sets that contain both hypotheses. That is, if spin 2700 and spin 2701 are linked by an [^1H , ^{13}C]-correlation peak, then the hypothesis set “spin 2700 \rightarrow ASN 46 H_α ” will not exist in the priority matrix. Instead, the hypothesis set will always contain at least two hypotheses, such as “spin 2700 \rightarrow ASN 46 H_α ” and “spin 2701 \rightarrow ASN 46 C_α ,” or “spin 2262 \rightarrow ASN 46 $\text{H}_{\beta 2}$ ” and “spin 2263 \rightarrow ASN 46 C_β ,” etc. By encoding heteronuclear-linked spins in such a manner, SideLink is forced to either accept them as matching covalently linked atoms, or not accept them at all, since all hypotheses of the set must be accepted or rejected together.

Redundant hypotheses must also be used to solve the inner combinatorial problem. Part of the reason for this becomes clear in consideration of the fact that a single spin, in some cases, can be assigned to more than one atom. For example, residues with two H_β s often only show one H_β resonance frequency due to rapid conformational averaging. In this case, the single H_β frequency observed must be assigned to both H_β s of the spin system. However, it must also be considered, that the H_β s may have different frequencies. Thus, the inner combinatorial problem must be encoded in such a way as to allow the assignment of one spin to more than one atom, if that is optimal, or alternatively to allow the assignment of the same spin to only a subset of those atoms if there are other spins to assign to the other atoms. To accomplish this, multiple hypothesis sets are created for each spin, each one allowing the acceptance of the spins match to a subset of the possible atoms to which it might fit. For example, if spin 1 might match both H_β s of a particular spin system (say ASN 46), then three hypothesis sets are created, “spin 2262 \rightarrow $\text{H}_{\beta 2}$,” “spin 2262 \rightarrow $\text{H}_{\beta 3}$,” and a third set which contains both hypotheses. This third hypothesis set allows SideLink to assign spin 2262 to both H_β s, while the first two hypothesis sets allow a singular match to either H_β should another spin matching the other H_β exist in the spin system.

In addition to the encoding of redundant hypothesis, functionally equivalent hypotheses must also be included in the priority matrix to solve the inner combinatorial

problem. This is in part because some atoms can give rise to cross-peaks that appear on different spectral lines. Consider for example an asparagine's H_{β} s. They may either appear in a ^{13}C -resolved NOESY along a single spectral line (if the local dynamics are fast on the NMR timescale), or in some cases they may be separated into two spectral lines. Prior to completing the side-chain assignments, it is impossible to know which case is true. If they do appear on separate spectral lines, however, their ^{13}C frequencies should be quite similar, since there is only one β carbon. Thus, hypotheses must be encoded into the priority matrix in a way that allows the assignment of separate spectral lines to the same carbon atom if the ^{13}C frequencies of the spectral lines are close together. At the same time, however, it must also be possible for the frequencies of the two separate spectral lines to be assigned to separate atoms of the spin system if that is optimal. These considerations are accomplished by creating three hypothesis sets in the priority matrix. Two of the hypothesis sets, e.g., "spin 2262 \rightarrow ASN 46 $H_{\beta 2}$ " and "spin 2263 \rightarrow ASN 46 C_{β} "; "spin 2268 \rightarrow ASN 46 $H_{\beta 2}$ " and "spin 2269 \rightarrow ASN 46 C_{β} ," allow separate competitive assignment of individual spins to ASN 46 C_{β} . However, if spin 2263 and spin 2269 are very close together in ppm, it must be considered that they may actually be from the same atom, so a third hypothesis is also included in the hypothesis set: "spin 2262 \rightarrow ASN 46 $H_{\beta 2}$ " and "spin 2263 \rightarrow ASN 46 C_{β} " and "spin 2268 \rightarrow ASN 46 $H_{\beta 3}$ " and "spin 2269 \rightarrow ASN 46 C_{β} ." By acceptance of this third hypothesis set, SideLink can effectively assign both spin 2263 and spin 2269 to the ASN 46 C_{β} .

There is another important use for redundant hypothesis encoding in the inner combinatorial problem, that of redundant peak discrimination. Often it is not possible before the side-chain assignment problem is solved to tell whether a peak in the [^1H , ^{13}C]-correlation projection of a 3D is a single peak or more than one peak overlapped. This is an especially relevant issue for automatic peak pickers, like the one used in the initial stage of SideLink to identify spectral lines of interest. Often the same spin system is picked as more than one spectral line of interest. This problem of redundantly-picked spectral lines is solved, however, in SideLink by the same mechanism that allows SideLink to assign two spins from distant spectral lines to the same atom described in the previous paragraph. In the case of overlapping spectral lines, when testing to see if the two lines can be added to the same spin system, a hypothesis set will be created that assigns the overlapping spins to the same atoms of the spin system, as well as individual hypothesis sets that allow separate atomic assignments to the overlapping spectral lines. Thus, if the optimal solution allows assigning the overlapping spectral lines to the same atoms, SideLink can do that by accepting the compound hypothesis. On the other hand, if separate assignments are optimal, the individual components of the compound hypothesis set also exist in the priority matrix and can be accepted separately whenever necessary. The ability of

SideLink to assign overlapping spectral lines to the same atoms in the inner combinatorial problem greatly reduces the dependence on the picking of correct spectral lines during the initial stages of the program's execution.

Once the inner combinatorial problem is encoded into a priority matrix, it can be solved by RHP logic. There is one further consideration that must be taken into account when evaluating spin system fitting. SideLink is not initially interested in whether a single assignment for each spin system is unambiguously true. Instead, since the fitting is only being used to generate a score measurement of whether an assignment could be true, the best fit combination of spin \rightarrow atom matches is used, irregardless of whether or not other reasonable possibilities exist. This is in contrast to the processing of the outer combinatorial problem where the only interesting solutions are those that can be determined with relative certainty. Since SideLink must make extensive use of spin-to-sequence matching, an acceleration mode has been incorporated into its RHP logic that allows it to bypass some of the complex encoding of the priority matrix for spin \rightarrow atom fitting, substituting some of the pre-RHP-processing elements for post-RHP processing elements in order to restore the solution to its full complexity. Post processing elements are sometimes advantageous since they do not have to be executed on hypothesis sets of the priority matrix that can be ruled out early in the RHP analysis.

From the accepted hypotheses of the inner combinatorial problem, a fitness score can be calculated as the geometric mean of the accepted spin \rightarrow atom priority scores. Only the best fitting spin for each atom and the best fitting atom for each spin is considered as part of the geometric mean. This geometric mean is the final goal of the inner combinatorial problem and is the overall "geometric fitness score" used to calculate the priority scores of the outer combinatorial problem.

2.8. Solving the outer combinatorial problem (spectral line grouping)

Once the spectral line comparison scores and the hypothetical sequence fit scores are known, the priority scores of the outer combinatorial problem can be calculated. Each score is calculated by the following equation:

$$\text{priority_score} = \text{cs}^a \times \text{sf}_{\text{spin}}^b \times \text{sf}_{\text{all}}^b \quad (4)$$

where cs is the score of the comparison spectral line of the spin to be grouped into the spin system with the amide spectral line of the spin system, sf_{spin} is the fit of the spin into the spin system at the best atom position available given the other spins of the system, sf_{all} is the geometric fitness score of all of all of the spins in the spin system, n is the number of spins fit into the system, and a and b are user-defined exponents to control the weighting of the relative terms. Both of the sequence fitting terms of the equation can optionally be turned off, as well, giving the user a wide

range of options as to how to incorporate the sequence fitting into the priority scores.

In addition to this equation, there are several exceptions that must be considered. First, since all hypotheses reflect the relative likelihood that a particular spin should be added to a particular spin system, the score of any hypothesis that requires the rejection of an already accepted hypothesis that has an otherwise better sequence fit score is reduced to 0. Additionally, if the acceptance of the new hypothesis would reduce the sequence fit score of such a hypothesis substantially (by more than [1-repeat bias] in most cases) then the score of the new hypothesis is also reduced to zero. The practical application of this is that no hypothesis is considered if the spins that are already in the spin system preclude the addition of the spins of the hypothesis.

Second, no hypothesis is considered if its acceptance would require the un-assignment of a spin which the program's user has designated prior to the program's execution. This allows the program to be used interactively as well as ensuring that its results must agree with the previously determined backbone resonance assignments.

Once the priority scores have been calculated (and therefore a priority matrix generated), the outer combinatorial problem can be evaluated by RHP. The result is a list of accepted hypothesis sets, each of which specifies the grouping of one spectral line with one spin system. Assignment of the individual spin frequencies of the spectral lines grouped into each spin system is trivial, as it can be accomplished as a single last iteration of the inner combinatorial problem executed once for each spin system.

2.9. Sample preparation and NMR spectroscopy

Two test molecules, doubly ^{13}C , ^{15}N -labeled CcmE [71,72] and doubly ^{13}C , ^{15}N -labeled mMjCM [73], were used to evaluate the program. Apo-CcmE is a 129 amino acid periplasmic domain of integral membrane cytochrom-*c* maturation heme chaperon E. The structure of the soluble domain of apo-CcmE shows two subdomains that are flexibly oriented relative to each other in solution. The structure of the N-terminal subdomain (residues I34–H130) displays high atomic precision, with a root mean square deviation (r.m.s.d.) of 0.6 Å for backbone heavy atoms constituting a well-defined core of the protein. This domain is formed by a six-stranded β -sheet wrapped to a closed β -barrel capped by an α -helix. The structurally less well defined C-terminal subdomain contains a single helical turn followed by an unstructured tail of 16 residues. In our studies CcmE represents a typical well structured compact protein. For the purpose of automatic side-chain assignment two NOESY spectra were selected, e.g., the ^{15}N -resolved NOESY data and the ^{13}C -resolved NOESY data for the aliphatics carbons acquired as a single [74] spectrum.

mMjCM is an artificially evolved monomeric chorismate mutase which adopts a molten globule state without ligand and that acquires native-like enzymatic activity by

'clamping-down' upon ligand (substrate) binding, demonstrating that rigid structure is not a prerequisite for efficient catalysis. In our laboratory we determined the 3D structure of mMjCM in a complex with its specific inhibitor (TSA) mimicking a putative transition state of natural ligand. Its 3D structure consists of an 4 α -helical bundle retaining a substantial degree of intramolecular mobility in millisecond time scale even in the complex with TSA (Vamvaca et al., paper in preparation). This protein represents flexible and partially dynamically disordered protein, which is considered as a difficult case in NMR work. For automatic assignment 4 3D NOESY spectra, e.g., backbone ^{15}N -resolved NOESY, arginine side-chain ^{15}N -resolved NOESY, aliphatic and aromatic ^{13}C -resolved NOESYs were used.

For both proteins reference resonance assignment, with which automatic resonance assignment was compared, was performed manually during the structure reconstruction process.

2.10. Calculations

Tests were run on one of two workstations: one with a 1.21 GHz Athlon processor and 512 Mb RAM and one with a 2.7 GHz Athlon 64 3700+ processor with 1 G RAM. The Athlon 64 processor was important for the aliphatic side-chain assignments as it brought the computation time down from approximately two days to ~4 h.

3. Results

SideLink has been tested on data acquired on two molecules of interest to our laboratory, mMjCM and CcmE. In each case we included only NOESY data and chemical shifts obtained during standard backbone assignment procedures for CcmE the amide and C_α chemical shifts were included, while for mMjCM additionally some of the C_β were also included (this was done just to test the program's compatibility with user assignments). In the case of mMjCM, there were three NOESY spectra acquired, a ^{15}N -resolved NOESY, and two ^{13}C -resolved NOESY s, one for optimized for aliphatic and one for aromatic groups. For CcmE, two NOESY spectra were acquired, as the ^{15}N -resolved NOESY data and the ^{13}C -resolved NOESY data for the aliphatics were acquired as a single [74] spectrum.

For each test case, the results of the initial identification of spectral lines gave comparable results. Like most peak-picking algorithms, SideLink's spectral line identification procedure picked several spectral lines that were from spectral artifacts, especially near the carrier and water frequencies. Additionally, there were many cases where several spectral lines were identified corresponding to fewer observed peaks in a [^1H , ^{13}C]-correlation spectrum. Though both of these complications could be addressed by increasing the stringency of the various parameters controlling the spectral line identification, instead the question-

able spectral lines were simply included in the RHP phase of the analysis without editing. This is analogous to the typical methodology used by human spectroscopists when they use programs to automatically pick peaks in spectra and later edit the peak list while assigning.

Despite the inclusion of extra spectral lines, SideLink performed well in assigning the side-chains for both of the test cases, averaging >75% of the side-chain C–H atoms assigned for the residues of the proteins where the backbone assignments were known. In the case of mMjCM, only 79% of the backbone assignments were known and ~20% of the corresponding spin systems had extremely low peak density in the NOESY spectra. Despite the limited data quality, 63% of the side-chain C–H atoms of the assigned residues were obtained with the remaining ones either not detected in the spectra or not unambiguously assignable.

For CcmE, a greater percentage, 91%, of the backbone amide resonances had been assigned, and SideLink was able to assign ~80% of the associated side-chain C–H resonances. Of particular interest, since the aliphatic ^{13}C -resolved NOESY data and the ^{15}N -resolved NOESY data were obtained simultaneously, in fact 75% of the C–H side-chain assignments were able to be obtained from a single NMR spectrum.

Because CcmE had been assigned prior to the development of SideLink, the program's results could be evaluated against the previously determined assignments. 93.4% of the aliphatic side-chain assignment groupings determined by SideLink either were consistent with the previous manually determined assignments or were assigned to atoms not previously assigned. The difference tolerance used to determine agreement was 0.07 for ^1H and 0.7 for ^{13}C and ^{15}N , which are the default difference thresholds used by the program to determine possible chemical shift equivalence. To do the comparison, obvious non-systematic assignment errors in the manual assignments were first excluded. These were identified as assignments that were more than 3 ppm away from their expected average values for protons, or more than 10 ppm away for ^{15}N and ^{13}C . 8 of the 36 differences were the result of the program assigning twin atoms (i.e., $\text{H}_{\beta 1}$ and $\text{H}_{\beta 2}$) separately rather than assuming they were overlapped as the spectroscopist had. Whether or not these eight are actually differences or simply new assignments not previously identified can only be determined by further analysis (structure calculation).

Only one significant deviation from the manually determined assignments was observed for the aromatic resonances, that for Y45 H_δ . This was caused by a large truncation artifact in the spectrum, which coincidentally made it look similar to the NOE pattern for Y45's amide proton. In fact, the quality of the ^{13}C -resolved NOESY used to observe the aromatic was poor enough that the manual assignments for the aromatics appear to have been done based on the NOE crosspeaks in the other NOESY spectrum rather than from the aromatic-optimized NOESY, as evidenced by the absence of any aromatic ^{13}C assignments in the manual assignments.

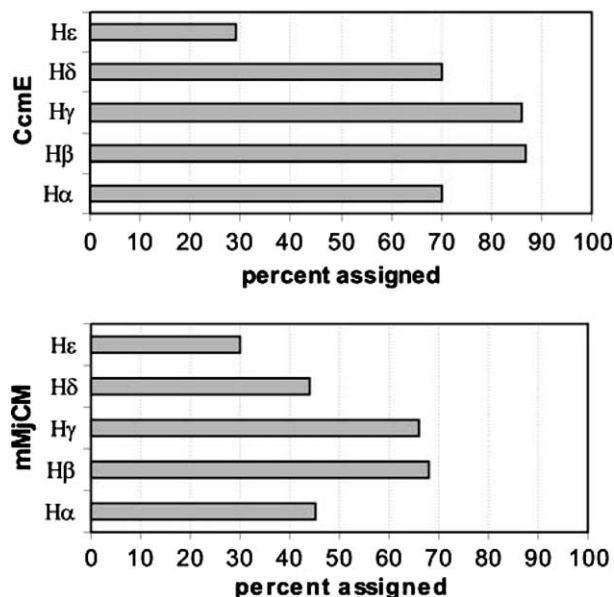


Fig. 8. Assignment distributions for SideLink for mMjCM and CcmE side-chain ^1H resonances. As expected, the percentage of assignments was affected by several factors including folding, spectral crowding, signal-to-noise ratio, spectral artifacts, and adjacency to the amide. In general, those nuclei closer to the amide were able to be assigned with a higher frequency than those more distant. This is partially due to the fact that the comparison score to the amide spectral line is one of the major criteria for assigning. The distal nuclei also tend to be more overlapped, however, and this also has a large affect on their assignability. Interestingly, a lower percentage of the H_δ s were assigned than for the C_{β} s despite their proximity to the amide and their relatively good resolution. This is due to the fact that most of them were folded in the spectra, resulting in lower signal-to-noise, and for many of them the signal-to-noise ratio was further reduced (or eliminated in several cases) by proximity to the $^1\text{H}_2\text{O}$ resonance frequency.

None of the assignments for the side-chain H–N resonances conflict with the manual assignments.

For both test molecules there was a clearly uneven distribution of assignments obtained with the percent of assignments for H_{β} s > H_{γ} s > H_{δ} s > H_{ϵ} (see Fig. 8).

4. Discussion

Recently many attempts have been made to assign all (or just backbone) resonances of proteins largely using chemical shifts (CAMRA [25]), NOESY (*St2nmr* [21], NOESY JIGSAW [24], CLOUDS [30], FIRE [75]) and residual dipolar couplings (NVR, [22,23], Tian et al. [26], and ABACUS [27]) supported just by a minimal set of through-bond correlation spectra. The performance of these approaches can be drastically improved by availability of 3D structures (or at least folds or secondary structures) of proteins. Here we explore the possibility of using just NOESY spectra in order to assign as many side-chain resonances as possible, providing a basis for automated structure reconstruction [1–3,5,6,76–78] Specifically, we discuss the potential merits and drawbacks of connecting automated resonance assignment with rapid structure (or fold) determination, which, in turn, is expect-

ed to enhance assignability of side-chains and provide a consistency check for the correctness of assignment.

Currently SideLink does not use structure calculation as a criterion for assignment, relying on its RHP logic to achieve a high level of success. Like SideLink, AutoLink also does not make use of structure calculation for making its conclusions. AutoLink, however, can incorporate at least some structural information from secondary structure, both in its chemical shift matching functions and in its medium-range NOEs use. This, however, still could be improved by the inclusion of long-range structural data which can only be obtained through structure calculation. Some previously developed programs use 3D structure data to increase their success rate (e.g., GARANT and ABACUS). Though these programs were based on genetic and Monte-Carlo algorithms, structure calculation was used as an effective tool for reducing the non-monotonic elements of the problem. While the RHP algorithm used by AutoLink and SideLink can directly address the non-monotonic NMR assignment problems, there is no restriction in the algorithm as to what factors can be included in the calculations. Thus, both AutoLink and SideLink can theoretically benefit from iterative execution, alternating with interleaved structure calculation, fold calculation, or secondary structure prediction procedures. Incorporation of 3D structural information into the RHP-driven process is relatively straightforward, as the only modification necessary to either program is that the priority scores for each process must be modified to take expected peak densities into consideration. Structural data can also be used to aid in picking peaks and assigning spin frequencies in the NOEs dimensions of spectra. Thus, it is theoretically optimal to construct an overall NMR structure determination outer loop that contains algorithms to address all four major components of NMR structure determination (backbone resonance assignment, side-chain resonance assignment, NOEs assignment, and structure calculation).

Though looping through structure calculation has significant benefit on resonance and NOESY assignment, it is risky to rely too heavily on it, since it is essentially circular reasoning. Errors in the upstream processes cause erroneous results in the downstream process, which then can reinforce the errors in the upstream processes leading to incorrect structure determination. Using downstream processes' results to reassess upstream process results is more reliable if the upstream processes can inform the downstream processes which part of their results are more or less reliable.

Both SideLink and AutoLink have the ability to assess relative certainty as a key component of their RHP processes. This is a key advantage for subsequent NOESY assignment and structure calculation. Though the programs' accuracies are high, the actual percentage of correct assignments is not nearly as critical as the relative certainty ratings. These certainty ratings can be used as weighting

factors for distance restraints derived from the chemical shift assignments and the NOESY data. Thus, relatively uncertain assignments that lead to relatively low-weighted restraints will have little effect on the structures calculated from them.

This is true for the currently available structure calculation methods, but even more so if RHP-based structure calculation methods are developed. RHP can be used both for NOESY assignment and structure calculation. In this case, inconsistent assignments would lead to inconsistent distance restraints, which would generally have no effect on the final structure because they would be ruled out as impossible by the RHP process before they affected any atomic coordinates. Since the structures would be based only on the internally consistent interpretations of the NMR data, they can then be used reliably as feedback to refine resonance assignments.

It is not necessary that all four of the overall structure determination loop's modules function by RHP in order to achieve a high level of reliability, as can be seen from the highly successful program CYANA [76] for NOESY assignment and structure calculation. It is necessary, however, that all four modules be able to incorporate confidence estimates from the other three modules into their calculations and be able to report the relative reliability of their own results. The effect of relative confidence communication between upstream and downstream modules, and vice versa, is that the uncertainty of the individual modules becomes linked into an overall uncertainty for the overall process. This prevents the creation of local minima in the solution space for the overall process based on the truncation of uncertainty information at the borders of the individual components of the outer level process.

Like AutoLink and SideLink, all components of the NMR structure determination process must be able to handle the non-monotonic nature of their functions in order to maintain the robustness of the overall process. A simple example of the robustness characteristic of a fully non-monotonic analytical process is the ability of SideLink to work with incorrectly picked spectral lines. The same factors considered by human spectroscopists in evaluating the results of a peak-picking program are considered by SideLink. Like a human spectroscopist, the program can assess the assignability of a particular spectral line in consideration of available atoms, similarity to other spectral lines, chemical shifts of the spectral line, the local noise level, and virtually all other relevant factors. Since all of the relevant factors are incorporated into a single decision process, grouping spectral lines (the outer combinatorial problem), early decisions made with only a subset of the relevant factors is avoided. This is important since, as with a human spectroscopist, use of fewer considerations can lead to unreliable results.

The ability of both SideLink and AutoLink to handle multiple factors simultaneously and function as numerical optimizers while avoiding local minima in solution space

is an inherent property of the RHP algorithm at their core, and is its main advantage over combinatorial global search and heuristic first-best approaches of the past. These other methods primarily focus on the analysis of solution space, using a logical energy function to assess the validity of any given state. Unfortunately, the NMR assignment problems are so non-monotonic, that there are always multiple local minima in the solution space and, since these minima are generally a function of several (often hundreds) of data elements, there is no easy way to avoid them in search of the global minimum.

One way to avoid local minima is to simply increase the amount of independent input data until only one minimum exists, that of the correct solution. The primary method used to increase the input data has been to take structure calculation results in consideration while assigning resonances. While this may seem circular (as described above), structure calculation includes much more information than simply the NMR data in terms of angle restrictions, Van der Waals forces, bond length restrictions, etc., and thus does include a large amount of data-independent elements that can aid resonance assignment. In fact, the structure calculation itself may be viewed as a simple method to format these data into terms that can more easily be used to restrict resonance assignments.

Other approaches to including more data have focused on narrowing the expected chemical shift ranges for sequence matching (MARS, [45], PISTACHIO [10]). For example, the program PISTACHIO uses coordinated spin system triplet matching where the expected chemical shift values are modified by using correlation statistics between related spins in order to reduce the allowed range for chemical shift targets. In contrast, most other programs use only single spin system for sequence comparison [81], and the expected chemical shift values for each component of the spin system are obtained from independent empirical averages. The more advanced sequence matching in PISTACHIO does allow the program to achieve a relatively high assignment accuracy (>90%), but still is not sufficient to prevent all errors.

MARS actually uses structural data (secondary structure prediction) to reduce the expected chemical shift ranges for spin systems. As with PISTACHIO, this increases the program's accuracy, but only in as much as the structural data is accurate.

The main drawback to using increased data to avoid local minima in solution space is that, for problems of variable size and complexity, there can never actually be a guarantee that there are no local minima despite the increase in input. Furthermore, increasing the size of the problem inevitably eventually overcomes the increases in data and, thus, such methods always have a limit to their accuracy that is a function of the size of the system they are applied to.

RHP analysis, on the other hand, doesn't actually focus on solution space at all, but rather on "priority space" which is a conjugate of the solution space and the "unique-

ness" of the solution components. Priority space is not complicated by multiple local minima, since it is logically impossible for likely, but incompatible solution components to be unique. Thus, RHP analysis focuses first to solve the relatively unambiguous parts of the problem and subsequently proceeds to the more ambiguous components. Thus, local minima are avoided because they must contain at least some ambiguous components in order to be local minima.

Noise is also a significant factor in the analysis of real data, since noise in data translates directly to noise in the solution space. Optimization methods that primarily rely on minimization of solution energy are therefore prone to noise-related errors. Noise in priority space, however, is greatly reduced, since by its nature noise is not unique. Still, priority space does contain some noise propagated from the data, and this is what necessitates the inclusion of repeat bias in the analysis (see Section 2.4). Essentially repeat bias determines a "uniqueness threshold" which any component of an optimized solution must exceed in order to be considered reliable. Other approaches to combinatorial problems may apply such a criterion after the optimization, but with RHP it is included during the solution process, preventing noise from trapping the solution search in a noise-created local minimum and also preventing noise from having a significant effect on the steps taken during the optimization process.

Like solution-space-based methods, RHP analysis is affected by the size of the problem it is applied to, but only because the complexity of the problem is related to the size. And, unlike solution-space-based methods, increasing the size (and complexity) of an assignment problem reduces the percent of returned assignments instead of reducing the accuracy of the assignments. Instead, accuracy is solely a function of the method of modeling the assignment problem, not the complexity or size of the problem.

However, as it was established by an analysis of optimization algorithms in NMR assignment applications [10] careful modeling of the assignment problem, adequate and full information content-retaining spectral representations are at the core of the success. Direct comparison of spectral lines instead of peak positions is a crucial element contributing to the robustness of SideLink. 3D ^{15}N - and ^{13}C -resolved NOESYs are parsed to a collection of 1D or 3D spectral lines along the indirectly detected ^1H dimension each designated by the root frequencies ^1H and X found in the 2D [^1H -X]-projections, where X is ^{15}N or ^{13}C , correspondingly. For the current application parsing (a representation of a 3D NOESY spectral matrix with a set of lower dimensionality objects, e.g., spectral lines) is done by direct extraction of spectral intensity vectors at designated positions in 3D spectra. Analysis shows that the primary limitation to SideLink's ability of assign spectra lies not in the RHP logic engine, but rather in the program's ability to process and compare spectral lines. It is in this area that most of the future improvements are expected to occur as better methods of spectral processing become

available, especially with regard to artifact recognition/suppression. One of the most promising spectral representation techniques is a unique multi-way decomposition of n D spectra as a collection of lower dimensionality objects, e.g., 1D spectral lines [79,80]. An ^{15}N -resolved NOESY spectrum can be decomposed into a sum of components, with each component corresponding to one or a group of peaks. Each component is defined as the direct product of three one-dimensional shapes. A consequence is reduction in dimensionality of the spectral data used in further analysis. These components can be used as a direct input to SideLink.

The division of the input data into functional components is a process requiring much care, however. While reduction of the 3D spectra to individual lines is of great aid in formatting and processing the assignment problem, it does, in fact, reduce the information content of the spectral lines somewhat. This is especially relevant for neighboring spectral lines where the peak width is larger than the separation between the lines. In this case, use of higher-dimensionality spectral slices can restore the lost data content and allow the separation of peak density into appropriate groupings. Alternatively, increasing the effective resolution of the spectra, either by decreasing the line-width or by increasing the number of resolving dimensions (i.e., 4D spectroscopy), can reduce the information loss upon separation of the data into spectral lines.

It should be re-iterated that, for our tests, SideLink required no user intervention whatsoever, except for input of the backbone assignments and the NMR spectra and setting its input parameters. Even the identification of folded chemical shifts was performed autonomously, with the program delaying the final determination of chemical shift folding into the assignment process whenever necessary.

As far as data is concerned, SideLink's algorithm has no specific requirements. The current development of the program, however, has focused on 3D spectra. Currently the main goal of NMR assignments is to obtain the ^1H resonances so that NOESY cross-peaks can be assigned. To obtain these, the user must provide at least one spectrum with intra-residue proton information. We have focused the development on the use of ^{13}C -correlated NOESY, but HCCH–TOCSY and HCCH–COSY are actually preferable (if they are available) since the interpretation of peak density is inherently less ambiguous for these spectra. Additionally SideLink can use any data available obtained from backbone assignment experiments. Though the user does need to provide the backbone assignments, individual cross-peaks from the backbone assignment spectra do not need to be assigned prior to execution. Such additional information, if provided, will be used by the program, allowing it to finish the assignments of an already partially assigned protein.

SideLink is not restricted to use on a single a [^1H , ^{13}C]-correlation spectrum, but can simultaneously consider the spectral lines from multiple spectra (e.g., ^{13}C -resolved NOESYs with different mixing times). Currently, we are

expanding the program's capabilities to make specific use of spectra obtained on partially deuterated proteins as well as residue-specific labeling, and use of spectra of variable dimensionality.

5. Conclusions

Though SideLink's development has focused on side-chain assignments, the direct spectral analysis approach used by the program can also be used to aid in backbone resonance assignment. Thus, it should be possible to expand our backbone assignment program (AutoLink) to work directly on spectra allowing backbone assignments to be obtained without prior peak (or spin system) picking by the user.

At the current stage of development, SideLink expects that the backbone resonances have been assigned prior to its execution. With minor modification of the program, however, it can be made to group spectral lines into spin systems whose backbone assignment is unknown, possibly to be determined later. Since such grouped resonances are useful in assigning backbone resonances, it should be possible to interface AutoLink and SideLink in such a way as to iteratively assign backbone resonances and side-chains coordinately. This may allow the assignment of at least some proteins with a reduced number of backbone assignment spectra, using ^{13}C -resolved NOESY and TOCSY data as a substitute.

As a final note, the RHP logic emulator designed for SideLink is not specific to NMR problems. It should also be possible to analyze other combinatorial problems as well using the same RHP engine.

Acknowledgment

Financial support was obtained from the Bruker Biospin Corporation through a grant to K.P.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.jmr.2006.03.012](https://doi.org/10.1016/j.jmr.2006.03.012).

References

- [1] C. Mumenthaler, P. Guntert, W. Braun, K. Wuthrich, Automated combined assignment of NOESY spectra and three-dimensional protein structure determination, *J. Biomol. NMR* 10 (1997) 351–362.
- [2] N. Oezguen, L. Adamian, Y. Xu, K. Rajarathnam, W. Braun, Automated assignment and 3D structure calculations using combinations of 2D homonuclear and 3D heteronuclear NOESY spectra, *J. Biomol. NMR* 22 (2002) 249–263.
- [3] Y. Xu, M.J. Jablonsky, P.L. Jackson, W. Braun, N.R. Krishna, Automatic assignment of NOESY cross peaks and determination of the protein structure of a new world scorpion neurotoxin using NOAH/DIAMOD, *J. Magn. Reson.* 148 (2001) 35–46.
- [4] M. Nilges, M.J. Macias, S.I. Odonoghue, H. Oschkinat, Automated NOESY interpretation with ambiguous distance restraints: the refined

- NMR solution structure of the pleckstrin homology domain from beta-spectrin, *J. Mol. Biol.* 269 (1997) 408–422.
- [5] T. Herrmann, P. Guntert, K. Wuthrich, Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS, *J. Biomol. NMR* 24 (2002) 171–189.
- [6] T. Herrmann, P. Guntert, K. Wuthrich, Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA, *J. Mol. Biol.* 319 (2002) 209–227.
- [7] A.J. Nederveen, J.F. Doreleijers, W. Vranken, Z. Miller, C. Spronk, S.B. Nabuurs, P. Guntert, M. Livny, J.L. Markley, M. Nilges, E.L. Ulrich, R. Kaptein, A. Bonvin, RECOORD: a recalculated coordinate database of 500+proteins from the PDB using restraints from the BioMagResBank, *Proteins* 59 (2005) 662–672.
- [8] W. Gronwald, H.R. Kalbitzer, Automated structure determination of proteins by NMR spectroscopy, *Prog. NMR Spectr.* 44 (2004) 33–96.
- [9] H.N.B. Moseley, G.T. Montelione, Automated analysis of NMR assignments and structures for proteins, *Curr. Opin. Struct. Biol.* 9 (1999) 635–642.
- [10] H.R. Eghbalnia, A. Bahrami, L.Y. Wang, A. Assadi, J.L. Markley, Probabilistic identification of spin systems and their assignments including coil-helix inference as output (PISTACHIO), *J. Biomol. NMR* 32 (2005) 219–233.
- [11] K.P. Neidig, M. Geyer, A. Gorler, C. Antz, R. Saffrich, W. Beneicke, H.R. Kalbitzer, Aurelia, a program for computer-aided analysis of multidimensional Nmr-spectra, *J. Biomol. NMR* 6 (1995) 255–270.
- [12] D.E. Zimmerman, C.A. Kulikowski, Y.P. Huang, W.Q. Feng, M. Tashiro, S. Shimotakahara, C.Y. Chien, R. Powers, G.T. Montelione, Automated analysis of protein NMR assignments using methods from artificial intelligence, *J. Mol. Biol.* 269 (1997) 592–610.
- [13] H.N.B. Moseley, D. Monleon, G.T. Montelione, Automatic determination of protein backbone resonance assignments from triple resonance nuclear magnetic resonance data, in: *Nuclear Magnetic Resonance of Biological Macromolecules*, Pt. B. 2001, Academic Press Inc: San Diego, pp. 91–108.
- [14] K.B. Li, B.C. Sanctuary, Automated resonance assignment of proteins using heteronuclear 3D NMR.2. Side chain and sequence-specific assignment, *J. Chem. Info.* 37 (1997) 467–477.
- [15] K.B. Li, B.C. Sanctuary, Automated resonance assignment of proteins using heteronuclear 3D NMR. Backbone spin systems extraction and creation of lypeptides, *J. Chem. Info.* 37 (1997) 359–366.
- [16] K.B. Li, B.C. Sanctuary, Automated extracting of amino acid spin systems in proteins using 3D HCCH-COSY/TOCSY spectroscopy and constrained partitioning algorithm (CPA), *J. Chem. Info.* 36 (1996) 585–593.
- [17] T. Szyperki, B. Banecki, D. Braun, R.W. Glaser, Sequential resonance assignment of medium-sized N-15/C-13-labeled proteins with projected 4D triple resonance NMR experiments, *J. Biomol. NMR* 11 (1998) 387–405.
- [18] B.E. Coggins, P. Zhou, PACES: protein sequential assignment by computer-assisted exhaustive search, *J. Biomol. NMR* 26(2003)93–111.
- [19] C. Bartels, P. Guntert, M. Billeter, K. Wuthrich, GARANT—a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra, *J. Comput. Chem.* 18 (1997) 139–149.
- [20] C. Bartels, M. Billeter, P. Guntert, K. Wuthrich, Automated sequence-specific NMR assignment of homologous proteins using the program GARANT, *J. Biomol. NMR* 7 (1996) 207–213.
- [21] P. Pristovsek, H. Ruterjans, R. Jerala, Semiautomatic sequence-specific assignment of proteins based on the tertiary structure—the program st2nmr, *J. Comput. Chem.* 23 (2002) 335–340.
- [22] C.J. Langmead, B.R. Donald, An expectation/maximization nuclear vector replacement algorithm for automated NMR resonance assignments, *J. Biomol. NMR* 29 (2004) 111–138.
- [23] C.J. Langmead, A. Yan, R. Lilien, L.C. Wang, B.R. Donald, A polynomial-time nuclear vector replacement algorithm for automated NMR resonance assignments, *J. Comput. Biol.* 11 (2004) 277–298.
- [24] C. Bailey-Kellogg, A. Widge, J.J. Kelley, M.J. Berardi, J.H. Bushweller, B.R. Donald, The NOESY JIGSAW: automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data, *J. Comput. Biol.* 7 (2000) 537–558.
- [25] W. Gronwald, L. Willard, T. Jellard, R.E. Boyko, K. Rajarathnam, D.S. Wishart, F.D. Sonnichsen, B.D. Sykes, CAMRA: chemical shift based computer aided protein NMR assignments, *J. Biomol. NMR* 12 (1998) 395–405.
- [26] F. Tian, H. Valafar, J.H. Prestegard, A dipolar coupling based strategy for simultaneous resonance assignment and structure determination of protein backbones, *J. Am. Chem. Soc.* 123 (2001) 11791–11796.
- [27] A. Grishaev, C.A. Steren, B. Wu, A. Pineda-Lucena, C. Arrowsmith, M. Llinas, ABACUS, a direct method for protein NMR structure computation via assembly of fragments, *Proteins* 61 (2005) 36–43.
- [28] A. Grishaev, M. Llinas, BACUS: a Bayesian protocol for the identification of protein NOESY spectra via unassigned spin systems, *J. Biomol. NMR* 28 (2004) 1–10.
- [29] A. Grishaev, M. Llinas, Protein structure elucidation from minimal NMR data: The CLOUDS approach, in: *Nuclear Magnetic Resonance of Biological Macromolecules*, Part C. 2005. pp. 261–295.
- [30] A. Grishaev, M. Llinas, CLOUDS, a protocol for deriving a molecular proton density via NMR, *Proc. Natl. Acad. Sci. USA* 99 (2002) 6707–6712.
- [31] R. Bernstein, C. Cieslar, A. Ross, H. Oschkinat, J. Freund, T.A. Holak, Computer-assisted assignment of multidimensional Nmr-spectra of proteins—application to 3d Noesy-Hmqc and Tocsy-Hmqc spectra, *J. Biomol. NMR* 3 (1993) 245–251.
- [32] N. Morelle, B. Brutscher, J.P. Simorre, D. Marion, Computer assignment of the backbone resonances of labeled proteins using 2-dimensional correlation experiments, *J. Biomol. NMR* 5 (1995) 154–160.
- [33] T.K. Hitchens, J.A. Lukin, Y.P. Zhan, S.A. McCallum, G.S. Rule, MONTE: an automated Monte Carlo based approach to nuclear magnetic resonance assignment of proteins, *J. Biomol. NMR* 25 (2003) 1–9.
- [34] N.E.G. Buchler, E.R.P. Zuiderweg, H. Wang, R.A. Goldstein, Protein heteronuclear NMR assignments using mean-field simulated annealing, *J. Magn. Reson.* 125 (1997) 34–42.
- [35] M. Leutner, R.M. Gschwind, J. Liermann, C. Schwarz, G. Gemmecker, H. Kessler, Automated backbone assignment of labeled proteins using the threshold accepting algorithm, *J. Biomol. NMR* 11 (1998) 31–43.
- [36] W.Y. Choy, B.C. Sanctuary, G. Zhu, Using neural network predicted secondary structure information in automatic protein NMR assignment, *J. Chem. Info.* 37 (1997) 1086–1094.
- [37] H.N. Lin, K.P. Wu, J.M. Chang, T.Y. Sung, W.L. Hsu, GANA—a genetic algorithm for NMR backbone resonance assignment, *Nucleic Acids Res.* 33 (2005) 4593–4601.
- [38] H.S. Atreya, K.V.R. Chary, G. Govil, Automated NMR assignments of proteins for high throughput structure determination: TATAPRO II, *Curr. Sci.* 83 (2002) 1372–1376.
- [39] H.S. Atreya, S.C. Sahu, K.V.R. Chary, G. Govil, A tracked approach for automated NMR assignments in proteins (TATAPRO), *J. Biomol. NMR* 17 (2000) 125–136.
- [40] K.P. Wu, J.M. Chang, J.B. Chen, C.F. Chang, W.J. Wu, T.H. Huang, T.Y. Sung, W.L. Hsu, RIBRA—an error-tolerant algorithm for the NMR backbone assignment problem, in: *Research in Computational Molecular Biology*, Proceedings, Springer-Verlag Berlin, Berlin, 2005, pp. 103–117.
- [41] R. Wehrens, C. Lucasius, L. Buydens, G. Kateman, Hips, a hybrid self-adapting expert-system for nuclear-magnetic-resonance spectrum interpretation using genetic algorithms, *Anal. Chim. Acta* 277 (1993) 313–324.
- [42] R. Wehrens, C. Lucasius, L. Buydens, G. Kateman, Sequential assignment of 2d-Nmr spectra of proteins using genetic algorithms, *J. Chem. Info.* 33 (1993) 245–251.
- [43] R. Wehrens, L. Buydens, G. Kateman, Validation and refinement of expert systems—interpretation of Nmr-spectra as an application

- in analytical-chemistry, *Chemometr. Intell. Lab. Syst.* 12 (1991) 57–67.
- [44] J.B. Olson, J.L. Markley, 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 (1994) 385–410.
- [45] Y.S. Jung, M. Zweckstetter, Mars—robust automatic backbone assignment of proteins, *J. Biomol. NMR* 30 (2004) 11–23.
- [46] J.E. Masse, R. Keller, AutoLink: automated sequential resonance assignment of biopolymers from NMR data by relative-hypothesis-prioritization-based simulated logic, *J. Magn. Reson.* 174 (2005) 133–151.
- [47] J.L. Pons, M.A. Delsuc, RESCUE: an artificial neural network tool for the NMR spectral assignment of proteins, *J. Biomol. NMR* 15 (2001) 15–26.
- [48] D. Auguin, V. Catherinot, T.E. Malliavin, J.L. Pons, M.A. Delsuc, Superposition of chemical shifts in NMR spectra can be overcome to determine automatically the structure of a protein, *Spectroscopy* 17 (2003) 559–568.
- [49] A. Marin, T.E. Malliavin, P. Nicolas, M.A. Delsuc, From NMR chemical shifts to amino acid types: investigation of the predictive power carried by nuclei, *J. Biomol. NMR* 30 (2004) 47–60.
- [50] A. Bax, G.M. Clore, P.C. Driscoll, A.M. Gronenborn, M. Ikura, L.E. Kay, Practical aspects of proton carbon carbon proton 3-dimensional correlation spectroscopy of C-13-labeled proteins, *J. Magn. Reson.* 87 (1990) 620–627.
- [51] L.E. Kay, M. Ikura, R. Tschudin, A. Bax, 3-Dimensional triple-resonance Nmr-spectroscopy of isotopically enriched proteins, *J. Magn. Reson.* 89 (1990) 496–514.
- [52] G.M. Clore, A. Bax, P.C. Driscoll, P.T. Wingfield, A.M. Gronenborn, Assignment of the side-chain H-1 and C-13 resonances of interleukin-1-beta using double-resonance and triple-resonance heteronuclear 3-dimensional Nmr-spectroscopy, *Biochemistry* 29 (1990) 8172–8184.
- [53] C. Griesinger, G. Otting, K. Wuthrich, R.R. Ernst, Clean Tocsy for H-1 spin system-identification in macromolecules, *J. Am. Chem. Soc.* 110 (1988) 7870–7872.
- [54] L.E. Kay, M. Ikura, A. Bax, Proton proton correlation via carbon carbon couplings—a 3-dimensional Nmr approach for the assignment of aliphatic resonances in proteins labeled with C-13, *J. Am. Chem. Soc.* 112 (1990) 888–889.
- [55] G.T. Montelione, B.A. Lyons, S.D. Emerson, M. Tashiro, An efficient triple resonance experiment using C-13 isotropic mixing for determining sequence-specific resonance assignments of isotopically-enriched proteins, *J. Am. Chem. Soc.* 114 (1992) 10974–10975.
- [56] B.A. Lyons, M. Tashiro, L. Cedergren, B. Nilsson, G.T. Montelione, An improved strategy for determining resonance assignments for isotopically enriched proteins and its application to an engineered domain of staphylococcal protein-a, *Biochemistry* 32 (1993) 7839–7845.
- [57] B.A. Lyons, G.T. Montelione, An Hcnh triple-resonance experiment using C-13 isotropic mixing for correlating backbone amide and side-chain aliphatic resonances in isotopically enriched proteins, *J. Magn. Reson. Ser. B* 101 (1993) 206–209.
- [58] G. Bodenhausen, G. Wagner, M. Rance, O.W. Sorensen, K. Wuthrich, R.R. Ernst, Longitudinal 2-spin order in 2d exchange spectroscopy (Noesy), *J. Magn. Reson.* 59 (1984) 542–550.
- [59] P.C. Driscoll, G.M. Clore, D. Marion, P.T. Wingfield, A.M. Gronenborn, Complete resonance assignment for the polypeptide backbone of interleukin 1 beta using three-dimensional heteronuclear NMR spectroscopy, *Biochemistry* 29 (1990) 3542–3556.
- [60] W.J. Fairbrother, A.G. Palmer, M. Rance, J. Reizer, M.H. Saier, P.E. Wright, Assignment of the aliphatic H-1 and C-13 resonances of the Bacillus-Subtilis glucose permease-Iia domain using double-resonance and triple-resonance heteronuclear 3-dimensional Nmr- spectroscopy, *Biochemistry* 31 (1992) 4413–4425.
- [61] S. Grzesiek, J. Anglister, A. Bax, Correlation of backbone amide and aliphatic side-chain resonances in C-13/N-15-enriched proteins by isotropic mixing of C-13 magnetization, *J. Magn. Reson. Ser. B* 101 (1993) 114–119.
- [62] B.T. Farmer, R.A. Venters, Assignment of side-chain C-13 resonances in perdeuterated proteins, *J. Am. Chem. Soc.* 117 (1995) 4187–4188.
- [63] A. Eletsky, O. Moreira, H. Kovacs, K. Pervushin, A novel strategy for the assignment of side-chain resonances in completely deuterated large proteins using C-13 spectroscopy, *J. Biomol. NMR* 26 (2003) 167–179.
- [64] L. Muller, Sensitivity enhanced detection of weak nuclei using heteronuclear multiple quantum coherence, *J. Am. Chem. Soc.* 101 (1979) 4481–4484.
- [65] G. Bodenhausen, D.J. Ruben, Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy, *Chem. Phys. Lett.* 69 (1980) 185–189.
- [66] R. Keller, The computer aided resonance assignment tutorial, CANTINA Verlag, Zurich, 2004.
- [67] R. Keller, Optimizing the process of NMR spectrum analysis and computer aided resonance assignment, in: *Biology. 2005, Diss. ETH Zurich NO. 15947 Zurich.*
- [68] J. Cavanagh, W.J. Fairbrother, A.G. Palmer, N.J. Skelton, *Protein NMR spectroscopy: principles and practice*, Academic Press, New York, 1996.
- [69] G.L.D. Boole, *An Investigation of the Laws of Thought*, MacMillon and Co, Cambridge, 1854.
- [70] P.J. Hurley, *A Concise Introduction To Logic*, Wadsworth Pub Co, Washington, 2005.
- [71] E. Enggist, L. Thony-Meyer, P. Guntert, K. Pervushin, NMR structure of the heme chaperone CcmE reveals a novel functional motif, *Structure* 10 (2002) 1551–1557.
- [72] F. Arnesano, L. Banci, P.D. Barker, I. Bertini, A. Rosato, X.C. Su, M.S. Viezzoli, Solution structure and characterization of the heme chaperone CcmE, *Biochemistry* 41 (2002) 13587–13594.
- [73] K. Vamvaca, B. Vogeli, P. Kast, K. Pervushin, D. Hilvert, An enzymatic molten globule: efficient coupling of folding and catalysis, *Proc. Natl. Acad. Sci. USA* 101 (2004) 12860–12864.
- [74] M. Sattler, M. Maurer, J. Schleucher, C. Griesinger, A simultaneous N-15,H-1-Hsqc and C-13,H-1-Hsqc with sensitivity enhancement and a heteronuclear gradient-echo, *J. Biomol. NMR* 5 (1995) 97–102.
- [75] T.E. Malliavin, P. Barthe, M.A. Delsuc, FIRE: predicting the spatial proximity of protein residues from 3D NOESY-HSQC, *Theor. Chem. Acc.* 106 (2001) 91–97.
- [76] P. Guntert, C. Mumenthaler, K. Wuthrich, Torsion angle dynamics for NMR structure calculation with the new program DYANA, *J. Mol. Biol.* 273 (1997) 283–298.
- [77] J.P. Linge, M. Habeck, W. Rieping, M. Nilges, ARIA: automated NOE assignment and NMR structure calculation, *Bioinformatics* 19 (2003) 315–316.
- [78] W. Gronwald, S. Moussa, R. Elsner, A. Jung, B. Ganslmeier, J. Trenner, W. Kremer, K.P. Neidig, H.R. Kalbitzer, Automated assignment of NOESY NMR spectra using a knowledge based method (KNOWNOE), *J. Biomol. NMR* 23 (2002) 271–287.
- [79] D.M. Korzhnev, I.V. Ibraghimov, M. Billeter, V.Y. Orekhov, MUNIN: application of three-way decomposition to the analysis of heteronuclear NMR relaxation data, *J. Biomol. NMR* 21 (2001) 263–268.
- [80] V.Y. Orekhov, I.V. Ibraghimov, M. Billeter, MUNIN: a new approach to multi-dimensional NMR spectra interpretation, *J. Biomol. NMR* 20 (2001) 49–60.
- [81] P. Guntert, M. Salzmann, D. Braun, K. Wuthrich, Sequence-specific NMR assignment of proteins by global fragment mapping with the program MAPPER, *J. Biomol. NMR* 18 (2000) 129–137.