Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Magnetic Resonance 174 (2005) 133–151
JMR Journal of Magnetic Resonance
www.elsevier.com/locate/jmr

AutoLink: Automated sequential resonance assignment of biopolymers from NMR data by relative-hypothesis-prioritization-based simulated logic

James E. Masse *, Rochus Keller

Department of Molecular Biology and Biophysics, ETH, Zurich 8039, Switzerland

Received 12 October 2004; revised 15 December 2004

Abstract

We have developed a new computer algorithm for determining the backbone resonance assignments for biopolymers. The approach we have taken, relative hypothesis prioritization, is implemented as a Lua program interfaced to the recently developed computer-aided resonance assignment (CARA) program. Our program can work with virtually any spectrum type, and is especially good with NOESY data. The results of the program are displayed in an easy-to-read, color-coded, graphic representation, allowing users to assess the quality of the results in minutes. Here we report the application of the program to two RNA recognition motifs of Apobec-1 Complementation Factor. The assignment of these domains demonstrates AutoLink's ability to deliver accurate resonance assignments from very minimal data and with minimal user intervention.

Keywords: NMR spectroscopy; Automation; Resonance assignment; Simulated logic; Relative hypothesis prioritization

1. Introduction

With the dawn of structural genomics, researchers worldwide have embarked on an ambitious new goal in biotechnology: to obtain at least one structure representing every protein fold [1]. This is reminiscent of the early stages of the genomics projects [2] where the goal had been defined long before the techniques had been developed to accomplish the task at hand. Indeed, the drive for more macromolecular structure-based knowledge has already lead to the development of several new methodologies, including rapid high-throughput micro-array crystal screening for crystallography, and rapid data collection methods like GFT NMR [3]. Prerequisite to the success of modern proteomics is the development of high-throughput data analysis and structure determination techniques.

NMR structure refinement has benefited from advanced semi-automated and automated NOE assignment and structure calculation software like ARIA [4] and CANDID [5,6]. Unlike for crystallographic data, however, the direct analysis of NMR data by computers has been largely limited due to the nature of the data. Crystallographic data consist of a regular array of points, whose relative intensity contains the structural data. NMR data, on the other hand, consist of a collection of non-discrete data points (peaks in spectra), which are initially used to determine the resonance frequencies of the various nuclei of the various spin systems contained within the macromolecule. The points of interest in an NMR spectrum are not constrained to a regular array, and often much of the data is simply not observable due to physical limitations of

* Corresponding author. Fax: +41 01 633 1294.
E-mail address: jmasse@phys.chem.ethz.ch (J.E. Masse).

1090-7807/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmr.2005.01.017

NMR spectroscopy. Furthermore, spectral data are often degenerate, leading to more than one possible solution unless several different types of NMR experiments are used simultaneously. For these reasons, automation of the initial stages of NMR data analysis, specifically assignment of resonance frequencies, has been problematic.

Earlier approaches to automated resonance assignment have generally required a high number of specialized NMR experiments [7]. The acquisition of these data is time consuming, and often not feasible, especially for larger molecules where dynamic limitations of the molecules hinder spectral acquisition. These specialized spectra are often quite tedious for viewing and use by human researchers due to the limited amount of data present in each one, and also due to the large number of spectra that must be used to make resonance assignments. This makes it very difficult for researchers to interact with the software in order to guide its analysis.

More recent methods have focused on developing algorithms that allow computers to work in a manner that is similar in approach to existing semi-automated methodologies, only with machines doing some of the work that traditionally is done by human operators [8,9].

We have developed a new program that expands on this latter approach. Our program, AutoLink, analyzes data from conventional NMR assignment spectra. Since it relies on intra- and inter-residue data that are incorporated into the "spin-system-based" object model of its host program, CARA [10], AutoLink can work with data from virtually any spectrum type, with no specific types required. Since the program can use conventional assignment spectra, it is easy for the user to review and edit the program's results using AutoLink's graphical displays and CARA's user-intuitive interface. The primary advantages of AutoLink over other existing programs largely result from more sophisticated "fuzzy logic" [11–16] and "relative hypothesis prioritization" (RHP), a process we have developed which simulates very "human-like" logic with a computer's speed and accuracy. The RHP approach gives AutoLink the unique ability to assess what can be determined in each NMR assignment problem and avoid over-assigning sequences which cannot be unambiguously determined. This allows the program to give reliable results even from minimal data. It is also significant that AutoLink does not use any absolute criteria for its assignment strategy, relying entirely on relative criteria for the evaluation of chemical shift assignments.

In this paper we report the use of AutoLink to assign two RNA recognition motifs (RRMs) of Apobec-1 Complementation Factor (ACF) [17] using a low number of NMR experiments.

2. Materials and methods

2.1. NMR samples

NMR samples of RRMs 2 (leucine 138-glycine 208) and 3 (leucine 233-glycine 293) from ACF were kindly provided by Christophe Maris in the laboratory of Frederic Allain (ETH, Zurich). The overall length of each protein construct was 108 and 115 residues for RRM 2 and 3, respectively, including sequences derived from the multiple cloning site of the pet22 expression vector. The RRM 3 sample was 2 mM protein, 200 mM NaCl, and 10 mM NaH_2PO_4, pH 6.5. The RRM 2 sample was 1 mM protein and 1 mM target RNA (sequence: UU**UGAUCAGUAUA**UCC—included for reasons of solubility), 10 mM NaCl, and 10 mM NaH_2PO_4, pH 6.5.

2.2. Spectroscopy

A^{15}N-HSQC [18–20], an HNCA [18–20], a CBCA (CO)NH [21], and a ^{15}N-NOESY-HMQC [18–20] were acquired for both RRMs 2 and 3 of ACF1. All spectra were acquired on a Bruker DRX 600 MHz spectrometer except for the NOESY spectra which were acquired on a Bruker 900 MHz spectrometer. See Table 1 for acquisition and processing parameters.

2.3. Computations

2.3.1. Algorithm overview

The main goal of AutoLink is to assign backbone resonances in macromolecules. For clarity, the majority of this discussion will assume that the molecule under investigation is a protein. Technically, AutoLink's algorithm can be used on any modular polymer for which NMR data are acquirable. This includes DNA and RNA since AutoLink is capable of using NOESY spectra.

In order to facilitate description of the algorithm used by AutoLink, it is necessary to define the term "spin system" as it applies to the program. A "spin system," for the purposes of this algorithm, is defined as a group of coupled resonances, called "spins," visualized as crosspeaks in one or more NMR spectra. This definition is inherited from AutoLink's host program CARA. Since AutoLink cannot yet directly view the NMR spectra, the spin systems must be defined by the user prior to running the program. This is a simple task and can be accomplished either manually using CARA's user interface to view and annotate the NMR spectra, or semi-automatically using an automatic spin-system-based peak picker followed by user inspection/editing of the results.

Once the user has identified the spin systems in the NMR data, AutoLink can then try to figure out which

Table 1
Spectral acquisition and processing parameters

	Dimension	Time domain (points)	Sweep width (ppm)	Dwell time (μs)	Carrier (ppm)	Apodization (function, phase)	Size (points)
RRM 2							
HSQC	^{1}H	2048	14	59.6	4.7	sin, π/2	2048
	^{15}N	256	36	228.3	116.0	sin, π/2	512
HNCA	^{1}H	2048	14	59.6	4.7	sin, π/2	2048
	^{15}N	64	34	241.81	118.0	sin, π/2	256
	^{13}C	60	32	103.54	56.0	sin, π/2	256
CBCA(CO)NH	^{1}H	2048	14	71.4	4.7	sin, π/2	2048
	^{15}N	76	34	290.15	118.3	sin, π/2	256
	^{13}C	72	60	66.26	53.0	sin, π/2	256
^{15}N-NOESY	^{1}H	2048	11	50.4	4.7	sin, π/2	1024
	^{15}N	80	42	130.5	116.0	sin, π/2	128
	^{1}H	256	11	50.5	4.7	sin, π/2	512
RRM 3							
HSQC	^{1}H	2048	14	39.9	4.7	sin, π/2	2048
	^{15}N	256	36	152.3	118.7	sin, π/2	512
HNCA	^{1}H	2048	14	59.6	4.7	sin, π/2	2048
	^{15}N	64	34	241.8	118.0	sin, π/2	256
	^{13}C	60	32	103.54	56.0	sin, π/2	256
CBCA(CO)NH	^{1}H	2048	15	55.6	4.7	sin, π/2	2048
	^{15}N	76	34	241.8	116.0	sin, π/2	256
	^{13}C	96	60	55.22	45.0	sin, π/2	256
^{15}N-NOESY	^{1}H	2048	11	50.4	4.7	sin, π/2	1024
	^{15}N	72	42	130.5	116.0	sin, π/2	128
	^{1}H	256	11	50.5	4.7	sin, π/2	512

Acquisition parameters for spectra used to assign RRMs 2 and 3 from ACF. None of the spectra were acquired for more than 48 h.

spin systems are adjacent in the protein sequence and form a "link" between them. In the case of conventional 3D protein assignment spectra, this can be viewed as "linking" amide-correlated strips within the 3D spectra. A set of two or more linked spin systems are called a "fragment." Once fragments have grown long enough, the chemical shifts of their spin systems can be used to assign the fragment to specific residues of the protein. This is the ultimate goal of the program (see Fig. 1 for a diagram of AutoLink's algorithm).

The algorithm used by AutoLink to form spin system links can be divided into two main parts, (1) spin system pair scoring (upper right box of Fig. 1) followed by (2) link hypothesis evaluation/re-evaluation (lower right box of Fig. 1). Spin system pair scoring is basically a search for potential spin system links based on comparing relevant spins within the spin systems. In order to find all potential matches, all possible spin system pairings must be considered and a "fitness" score is calculated for each. A pair of spin systems and their score is called a "link hypothesis." The fitness score for any particular link hypothesis is actually itself a function of one or more sub-scores. The sub-scores are calculated by comparing specific resonances within the spin system pair (i.e., comparing the C_{α} spin of one spin system with the $C_{\alpha-1}$ spin of another spin system, comparing the C_{β} and $C_{\beta-1}$ spins of spin systems, comparing all of the NOE spins of one spin system with those of another spin system, etc.). The actual mathematical functions used to compare spins within spin systems, and how the individual sub-scores are combined in order to calculate the overall fitness score for each link hypothesis, are user-defined. A more detailed description of the evaluation of spin-system pair fitness is described in Section 2.3.2. Together all of the link hypotheses form a list of potential spin system links called the "priority list." Thus the priority list can be viewed as a set of link hypotheses, each associated with a fitness score, which is a function of spectral data only, and without any consideration of the protein sequence.

Once the priority list has been created, AutoLink starts "hypothesis evaluation/re-evaluation" cycles (Lower right box of Fig. 1). These cycles involve modifying the link hypothesis fitness scores by factors that are dependent on the acceptance or rejection of other link hypotheses.

The first modification of the link hypotheses fitness scores in each cycle involves considering how the resulting fragment formed by any given link hypothesis would fit into the protein sequence. These newly modified scores are stored in the "base priority prime list."

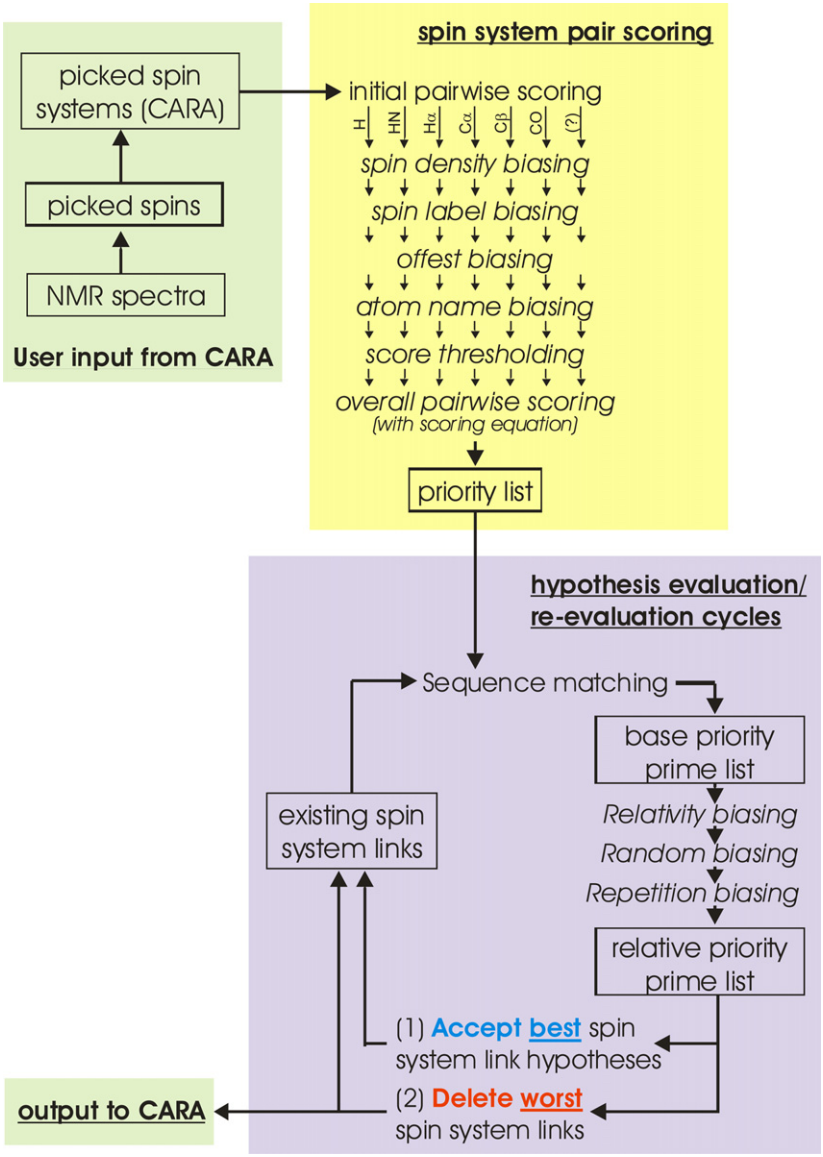


Fig. 1. Schematic showing overall flow of AutoLink algorithm. Input (upper left box) is obtained from CARA in the form of pre-defined spin systems consisting of correlated resonances in NMR data. Initially this input is used to generate a series of spin system link hypotheses which together make up the priority list (upper right box). The link hypotheses are composed of a spin system pair and a score representing a relative probability that the spin system pair corresponds to adjacent residues in the protein sequence. After the initial scoring is completed, the scores in the priority list are modified during the hypotheses evaluation/re-evaluation cycles (lower right box) according to matching of hypothetical fragments to the protein sequence, creating the base priority list, and subsequently by a series of logical biases, creating the relative priority prime list. These logic biases (''relativity biasing,'' ''random biasing,'' and ''repetition biasing'') are discussed in Sections 3.2.1–3.2.3, respectively, and are important in AutoLink's logical decision making. During these cycles AutoLink both accepts and rejects link hypotheses based on relative criteria (RHP). Once AutoLink has ''decided'' that no more link hypotheses are determinable, the program halts the hypothesis evaluation/re-evaluation cycles and reports its results to the user through its graphical interface (see Fig. 5) and output files.

During the initial hypothesis evaluation/re-evaluation cycle, of course, the entire protein sequence is yet unassigned, and is thus available for consideration for each link hypothesis. In later rounds, once some sequences of the protein have already been assigned, these prior assignments must be taken into account when evaluating link hypotheses. Details of how AutoLink fits hypothetical fragments to the protein sequence are presented in Sections 3 and 3.1.

After consideration of fragment fitting to the protein sequence, the base priority prime list is then further processed by logical biases, which are the heart of AutoLink's RHP decision-making process. These biases allow the program to mimic ''human'' logic and are

described in detail later in Section 3.2. The final, logically biased list is called the "relative priority prime list" and contains a series of link hypotheses, ranked according to their relative likelihood of being consistent with the NMR data, the protein sequence, and with other link hypotheses. Once the relative priority prime list has been created, AutoLink can make the decision to accept or reject link hypotheses based on the fitness scores in the list. However, since acceptance and rejection of link hypotheses affect the scores within both the base priority prime list and the relative priority prime list, it is only possible to accept or reject a small number of link hypotheses before these lists must be re-calculated. Thus, only a few link hypotheses are accepted or rejected during each cycle, and each cycle is followed by a re-calculation of the priority prime lists. The process ends when the program determines that none of the remaining link hypotheses can be accepted, and that there are no reasonable alternatives for the existing spin system links. Evaluation, acceptance, and rejection of link hypotheses are described in much greater detail in Section 3.2.4.

2.3.2. Spin system pair scoring

The initial stage of the algorithm loops through all spin system pairs and calculates scores that are dependent on the relatedness of the spins of the spin systems. These scores can be viewed as a measure of how likely each spin system pairing corresponds to adjacent spin systems within the protein with higher scores signifying higher probability.

Each spin system pair is scored according to one or more user-selected types, i.e., C_α, C_β, CO, H (NOE), etc. Each type selected generates a separate "sub-score" for any given spin system pair. The overall score for each spin system pair is then calculated as a user-defined combination of the individual sub-scores (see Section 2.3.2.6). A schematic of the overall approach to spin system pair scoring is shown in Fig. 2.

Calculation of each sub-score type is governed by the following equation:

$$\sum_{s1}\sum_{s2} f(\Delta). \tag{1}$$

Here $\sum_{s1}$ and $\sum_{s2}$ stand for all of the spins (resonances) comprising each spin system (spin system 1 and spin system 2, respectively), and $f(\Delta)$ is a user-defined function describing a mathematical comparison of each spin of spin system 1 with each spin of spin system 2 depending on their chemical shift difference and modified by user-controlled factors (described below). In general the relatedness of each spin of spin system 1 to each spin of spin system 2, $f(\Delta)$, is defined by the following equation:

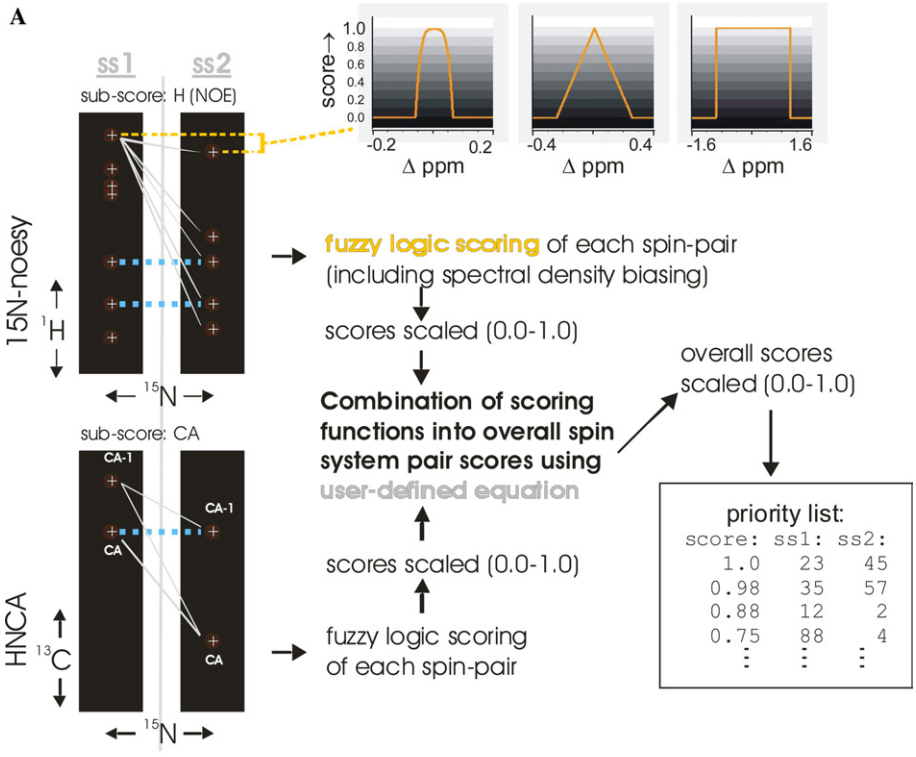
$$f(\Delta) = rel\_score = 1 - \frac{(a \times \Delta)^b}{a} \times c, \tag{2}$$

where a, b, and c are user-defined constants optimizable with AutoLink's graphical interface. Δ is the absolute value of the difference in the chemical shifts of the two spins. If *rel_score* is negative it is increased to 0. Thus, the basic spin pair scoring function produces values that range from 0 to 1, with spins that are further apart from each other in ppm scoring lower than those that are close together. Only spin pairs with labels or atom types appropriate to the score type are considered relevant. For example, when scoring C_α's, only spins that are considered by AutoLink to potentially be C_α's or $C_{\alpha-1}$'s are used to generate the spin system pair score. The identity of each spin is determined by either its label (i.e "C_α" or "$C_{\alpha-1}$") in CARA if it has been defined by the user, or by the atom type ("*C*" for carbon nuclei) in conjunction with chemical (i.e., $36 < \text{chemical shift} < 76$ ppm for compatibility with C_α scoring). Each spin pair relatedness score is additionally biased according to several user-controlled functions, which determine how the program interprets spin density, overlap, the labeling of each spin, the atom name in each atom label, and the offset in spin label. Each of these biases is described in detail below.

2.3.2.1. Spin density bias. Spins that occur at uncommon chemical shifts are often more valuable for determining spin system linkages than spins that occur at relatively common ppm values. For this reason it is useful to bias the spin scores accordingly. In AutoLink this is called spin density biasing and is controlled by a user-defined parameter which has valid settings between 0 and 1. Spin pair scores are modified according to the formula:

$$rel\_score' = (1 - sdc) \times rel\_score + \left(\frac{sdc \times score_{1\to 2}}{density_{1\to\Sigma}}\right), \tag{3}$$

where *rel_score* is the base relatedness score of spin 1 and spin 2 of the spin pair, *rel_score'* is the density-compensated relatedness score, and *sdc* is a user-defined input parameter. $density_{1\to\Sigma}$ refers to the average base relatedness score of spin 1 with all spins from all of the spin systems and is called the *spin density*. Thus the spin pair relatedness scores are adjusted proportionally to the inverse of spin density of the first of the spins in the spin pairs. This causes spins with more unusual chemical shifts to weigh more heavily than those with common chemical shifts when calculating the relatedness of spin systems. As with most of AutoLink's controls, spin density biasing can be partial ($0 < sdc < 1$), with only a fraction of the base score being adjusted, none ($sdc = 0$), or all ($sdc = 1$). Spin density biasing is generally useful for scoring of all data types, but is especially important for scoring high-density spins like ^{1}Hs in NOESYs.



B

equation:	interpretation:
Cα	Consider only spin system pairs with compatible Cα (ignore other spins).
H×Cα	Consider only spin system pairs with compatible Cαs and H xpeaks (NOESy) in common.
H×Cα×Cβ	Consider only spin system pairs with compatible Cαs, compatible Cβs, and H xpeaks in common.
Cα×(H+Cβ)	Consider only spin system pairs with compatible Cαs and either compatible Cβs or H xpeaks in common.
H&&Cα	Consider only spin system pairs with compatible Cαs and H xpeaks in common.
H||Cα	Consider only spin system pairs with either compatible Cαs or H xpeaks in common.
H&|Cα	Consider only spin system pairs with either compatible Cαs or H xpeaks in common, but especially favor those with both.
$C\alpha^2\times(H+C\beta+CO)$	Consider only spin system pairs with stringently compatible Cαs and at least one of the following: H xpeaks in common, compatible Cβs, or compatible COs.

Fig. 2. (A) Diagram of spin-system pair scoring. The spins of both spin systems (represented here by crosspeaks in spectra) are compared using a user-defined scoring function. Spin comparisons are signified by lines connecting relevant points in the NMR spectra. Model sub-scoring functions are plotted in gold at the top of the figure. The sub-scores are calculated by adding together all of the relevant comparison scores of each spin of one spin system (ss1) with each spin of another spin system (ss2). After the individual sub-score types (C_α and H (NOE) are shown here) are calculated, they are then combined to form an overall spin-system pair score. The manner in which they are combined is determined by a user-defined scoring equation (see Section 2.3.2.6 in the main text). The overall scores are recorded in the "priority list." (B) Examples of scoring equations and literal interpretation. The left column shows examples of valid scoring equations demonstrating a variety of operations including the quasi-operators "&&," "||," and "&|." The literal interpretations in the right column are pneumonic approximations of the operations presented in the scoring equations.

2.3.2.2. Assigned spin bias. When picking peaks, it is sometimes impossible to be sure what spin assignment to give to a particular spin. Therefore, AutoLink can work with unlabeled spins. It is, however, logical to bias scoring in favor of less ambiguous spins. This bias can be controlled in AutoLink according to the equation:

$$rel\_score' = (1 - asb) \times rel\_score_{unlabeled} + asb \times rel\_score_{labeled}, \quad (4)$$

where rel_score' is the modified score, asb is the user-defined input parameter, $rel\_score_{unlabeled}$ is the spin pair score irregardless of spin labeling, and $score_{labeled}$ is the spin pair score considering labeling. If both of the spins of the spin pair are labeled, then $score_{labeled} = score_{unlabeled}$. Otherwise $score_{labeled} = 0$. Thus if $asb = 1$ only spin pairs where both spins involved are labeled can score >0. Likewise, if $asb = 0$, all spin pair labels will not affect the spin pair score at all. Intermediate values for asb cause fractional biasing of the spin pair score. Biasing for assigned spins is important for most assignment spectra, where most of the spin labels are relatively easy to determine. It is generally useful to set $asb = 0$ when one needs to assign spin systems whose spins cannot be unambiguously labeled prior to backbone resonance assignment (such as ^{1}H spins obtained from NOESY spectra). Additionally, assigned spin bias can be used to help assign overlapped spin systems. In such cases, one can include each spin

in each of the overlapped spin systems, set *asb* <1, and allow AutoLink to try and find the best related spin systems in the absence of predetermined spin assignments. This is helpful because it allows AutoLink to search for possible spin system pairs without user-bias.

2.3.2.3. Offset bias. For some spins, while it may be possible to identify the type of nucleus (C_α, for example), it is sometimes difficult to determine whether the crosspeak represents an intra- or inter-residue correlation. This is fairly frequent, for example, when identifying C_α and $C_{\alpha-1}$ peaks in HNCA spectra in cases where no corroboration from an HN(CO)CA is possible either because no such spectrum is available or because the spin system in question does not appear in the HN(CO)CA spectrum. Occasionally the C_α and $C_{\alpha-1}$ crosspeaks are occasionally nearly equal in intensity in the HNCA or there is too little signal above the noise to be sure which is which. In order to address this potentiality, it is possible to decrease the final score's dependence on the offset of the spin label (i.e., the "−1" in "$C_{\alpha-1}$"). In AutoLink this is accomplished by modifying the $score_{labeled}$ component in the above equation as follows:

$$rel\_score' = (1 - ob) \times rel\_score_{labeled} + ob \times rel\_score_{labeled+offset}, \quad (5)$$

where $rel\_score'_{labeled}$ is the new labeled score component, *ob* is the user-defined control parameter, $rel\_score_{labeled}$ is the spin pair score considering the presence of labels but ignoring offsets, and $rel\_score_{labeled+offset}$ is the spin pair score considering the offset. If *ob* is 0, then the offsets in the spin labels will be irrelevant to the spin pair relatedness score. If *ob* is 1, then scores >0 are only possible if the two spins have compatible labels (i.e., spin 1 has offset 0 and spin 2 has offset −1). Of course, intermediate values for *ob* cause fractional offset biasing.

2.3.2.4. Atomic assignment bias. This bias works exactly as for offset biasing, except that the atomic assignment (i.e., the "C_α" in "$C_{\alpha-1}$") is the critical element rather than the offset:

$$rel\_score' = (1 - aab) \times rel\_score_{labeled} + aab \times rel\_score_{labeled+same}, \quad (6)$$

where $rel\_score'_{labeled}$ is the new labeled score component, *aab* is the control parameter, $rel\_score_{labeled}$ is the spin pair score considering the presence of labels but ignoring the atomic assignment, and $rel\_score_{labeled+same}$ is the spin pair score considering the atomic assignment. In general atom types are considered compatible if they are the same (i.e., "C_α" and "C_α," or "C_β" and "C_β").

2.3.2.5. Score threshold. After all spin system pairs have been scored, the scores for each sub-score type are linearly scaled to values between 0 and 1. Any spin system pair sub-score below a user-defined threshold is reduced to 0 and no further consideration of this spin pair is given for that sub-score type. This does not significantly change the scoring results, as in most cases an extremely low threshold is chosen, but is important for reasons related to computation time. For most proteins, since each residue is represented by one spin system, the number of spin system pairs can be as high as the number of protein residues squared (minus the self pairs, of course). For most sub-score types, however, most of the scores for the spin pairs are 0 or very close to 0. Thus, a threshold filter prevents the program from wasting computation time on useless calculations like multiplication of zeros by zeros in later stages of analysis.

2.3.2.6. Overall spin system pair scoring. Once each spin system pair has been scored in all of the desired score types, these individual scores are combined to form an overall score for each spin system pair (the spin system link hypothesis). These combined scores are then sorted according to their value and stored in the priority list. The priority list, thus, contains a list of spin system link hypotheses sorted by their overall fitness scores, which are a measure of how likely the spin systems in the pair are to be adjacent in the protein sequence.

The formula used to combine the individual spin pair sub-scores is a user-defined equation. See Fig. 2B for examples. All of the standard arithmetic operators are valid for this "scoring equation" (addition/subtraction, multiplication/division, and exponentiation). The equation editor in AutoLink is fully capable of interpreting nested parenthesis, so complex spectral equations can be defined. In practice, however, only simple equations are generally used since for any given protein only a few assignment spectra are usually obtained, and thus, there are only a few sub-score types available.

The most useful standard operators used in combining sub-score types are addition and multiplication, which can be viewed as analogous to a "fuzzy OR" function and a "fuzzy AND" function, respectively. In addition to the standard operators, AutoLink can interpret three non-standard operators, "‖," "&&," and "&|." The ‖ operator is defined as a "quasi-OR" operation. It is actually a shorthand for the average of its operands (i.e., $A \parallel B = (A + B) \times 0.5$). The && operator, called a "quasi-AND," is defined as the geometric average of its operands. Thus $A \,\&\&\, B = \sqrt{A \times B}$. The last operator, &|, or "quasi-AND/OR" is a combination of the other two quasi-operators defined as $A \,\&|\, B = (A \,\&\&\, B) \times 0.5 + (A \parallel B) \times 0.5$. These operators are remarkably useful for defining complex scoring functions because their product is in the same range as their operands, allowing easy combination of multiple scores without worrying about scaling.

Use of mathematical operators instead of true Boolean logic allows AutoLink to compare spin system pairs

quantitatively rather than merely qualitatively (see Table 2 for a comparison of Boolean logic operators with the "fuzzy" operators usable by AutoLink). In effect, AutoLink's comparison of spin systems is entirely fuzzy-logic based, starting from the initial spin–spin scores and propagating all uncertainties all the way through to the overall spin system pair score. Thus the program can avoid making decisions based on individual data points in favor of only making decisions based on the combined match of multiple data points at later steps in the analysis. Fuzzy logic is also useful because it allows AutoLink to bias in favor of spin system pairs that have exact matches in the sub-scoring functions over spin system pairs with more marginal matches. This is particularly useful in dealing with high-density spectra (like NOESYs) and reduces the dependence upon correct peak-picking and spin system identification by the user prior to running the program.

Table 2
Comparison of Boolean, "fuzzy," and "quasi" operators

Function name/equation	Operand 1 (x)	Operand 2 (y)	Result
Boolean AND n/a	1	1	1
	1	0	0
	0	0	0
	0.5	0.5	n/a
	1	0.5	n/a
Fuzzy AND $x \times y$ = product	1	1	1
	1	0	0
	0	0	0
	0.5	0.5	0.25
	1	0.5	0.5
Quasi-AND $\sqrt{x \times y}$ = geometric average	1	1	1
	1	0	0
	0	0	0
	0.5	0.5	0.5
	1	0.5	0.707
Boolean OR n/a	1	1	1
	1	0	1
	0	0	0
	0.5	0.5	n/a
	1	0.5	n/a
Fuzzy OR $x + y$ = sum	1	1	2
	1	0	1
	0	0	0
	0.5	0.5	1
	1	0.5	1.5
Quasi-OR $(x + y) \times 0.5$ = average	1	1	1
	1	0	0.5
	0	0	0
	0.5	0.5	0.5
	1	0.5	0.75
Quasi-AND/OR $\frac{\sqrt{x \times y} + (x+y) \times 0.5}{2}$	1	1	1
	1	0	0.25
	0	0	0
	0.5	0.5	0.5
	1	0.5	0.729

Demonstration of the function of Boolean operators with "fuzzy," and "quasi" operators used by AutoLink to compute overall spin system pair scoring. As signified by "n/a" in the table, Boolean operators are not applicable to values other than 0 and 1. The use of fuzzy operators allows AutoLink to propagate uncertainties in individual scoring functions into the overall combined result. As can be seen from the examples, the quasi-operators defined for AutoLink are similar in function to the fuzzy operators, but maintain the overall value range of the operands. The quasi-AND/OR operator performs an intermediate function between a quasi-AND and a quasi-OR.

After all of the individual scores are combined, the resulting overall scores are re-scaled to values between 0 and 1, and then filtered according to a user-defined threshold. This filter is important for the functionality of the program, because it removes all insignificant spin system pairs from the priority list. Since most spin system pairs are only coincidentally related and therefore score 0 or very near 0, filtering even with a very low threshold greatly shortens the base priority list and consequently reduces the number of calculations AutoLink must perform.

At this point the priority list has been created. It contains information about the relatedness of spin systems based on the spectral peaks alone. Since the scores in this list are not a function of other spin system links, the base priority list must only be calculated once at the start of the analysis. All of the processing of the link hypotheses past this point, however, is dependent on the acceptance or rejection of other spin system link hypotheses, and so all of this processing (i.e., calculation of the priority prime lists) must be repeated periodically during the "hypothesis re-evaluation cycles" described below.

3. Hypothesis evaluation/re-evaluation cycles

3.1. Calculation of the base priority prime list (sequence matching)

As stated above, the priority list undergoes modifications which are dependent upon the current link hypotheses that have already been accepted as well as existing spin system—residue assignments. These modifications produce first the "base priority prime list" and then subsequently the "relative priority prime list."

The base priority prime list is created from the priority list by considering how well the fragments that would be formed upon the acceptance of each hypothesis would match positions of the protein sequence:

$$\begin{aligned} fragmentScore = [&score_{rel}(ss_1, res_1) \\ &\times score_{rel}(ss_2, res_2) \times \cdots \\ &\times score_{rel}(ss_n, res_n)]^{1/n}, \end{aligned} \tag{7}$$

where *fragmentScore* is the overall score for the fragment, n is the number of spin systems in the fragment, $ss_1, \ldots, ss_n$ are the spin systems of the hypothetical fragment, res_1–res_n represent the residues of the protein at the relevant positions in the sequence, and

$score_{rel}(ss_x, res_y)$ is the *relative* score of spin system x against residue position y in the protein sequence when compared to its match to all other residue positions of the protein sequence.

In order to evaluate the relative score, $score_{rel}(ss_x, res_y)$, each spin system is first scored against each possible position of the protein sequence according to the following equation (also see Fig. 3):

$$score_{abs}(ss_x, res_y) = \left(\sum \left[s \times \left(1 - \left|\frac{chemShift_{obs} - chemShift_{resAvg}}{standev_{ref}}\right|\right)\right]\right) \Big/ \#spins, \tag{8}$$

where $\#spins$ is the number of possible matching spins between the spin system and the residue, $chemShift_{obs} - chemShift_{resAvg}$ is the difference in the chemical shift of a spin from the experimentally observed spin system from the empirically determined average chemical shift for that spin, $standev_{ref}$ is the empirically determined standard deviation of the chemical shift for the comparable spin of the residue the spin system is being matched to, and b is a user-defined slope of a line.

$$\sum \left[s \times \left(1 - \left|\frac{chemShift_{obs} - chemShift_{resAvg}}{standev_{ref}}\right|\right)\right],$$

therefore, is the sum of all of the individual scores from pairing each spin in the system being tested with all of the comparable average spins of the residue to which the spin system is being matched. The above equation is logically modified such that any term

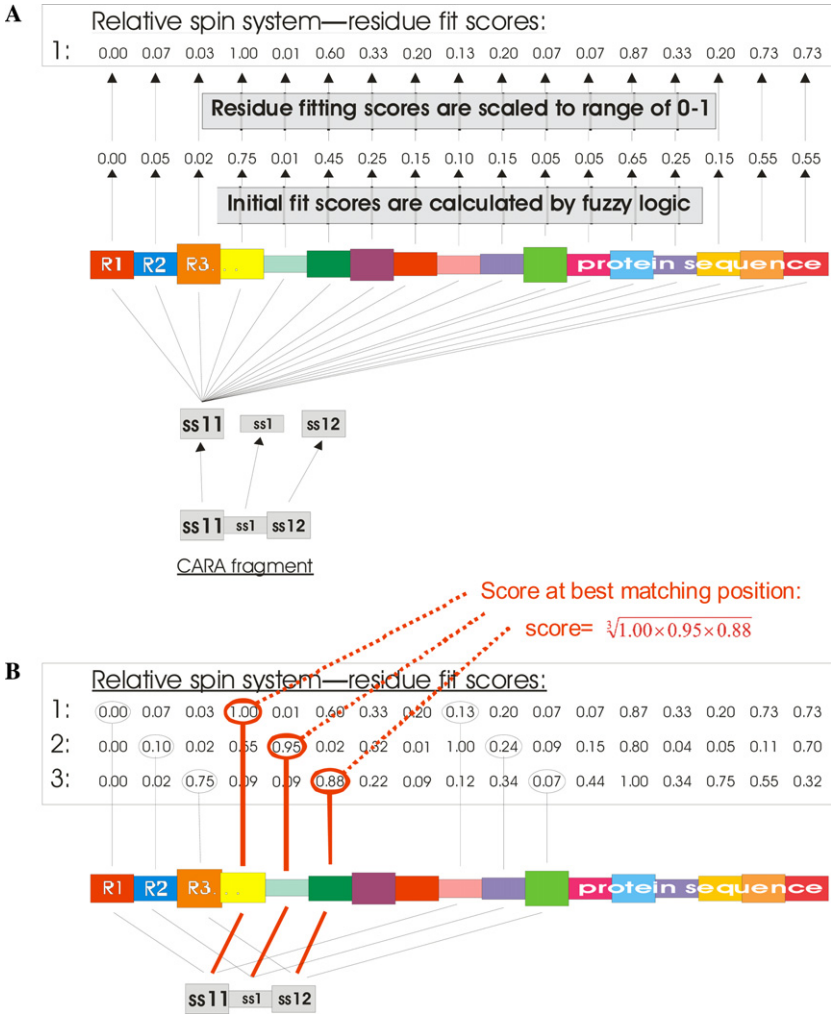


Fig. 3. Matching of spin system fragments to the protein sequence proceeds in two stages, (A) individual spin system—sequence matching, and (B) fragment—sequence matching. (A) Each spin system of each fragment is first fit using fuzzy logic to each residue position of the protein sequence (see Section 3.1 in the main text). In this diagram comparison of a single spin system (ss11) to each position of the protein sequence is signified by solid black lines. Though not shown here, ss1 and ss12 are also fit to each position of the protein sequence. These fit scores for each spin system are then scaled to a range of 0–1. The final fitting scores for each spin system to the protein sequence are shown here in rows 1:, 2:, and 3: for ss11, ss1, and ss12, respectively, in the "Relative spin system—residue fit scores" boxes. (B) Fragment—sequence fitting: Fragment scores are calculated as the geometric average of the single spin system relative fits for consecutive positions within the protein sequence. Sequence fitting also takes into account additional criteria such as already-assigned fragments and specific-isotopic-labeling data.

$$\left(1 - \left|\frac{chemShift_{obs} - chemShift_{resAvg}}{standev_{ref}}\right|\right)$$

which results in a negative value is increased to 0. Thus, the spins of any given spin system produce scores ranging from 0 to 1 which are added together to obtain a total score for the spin system. Since this score is then divided by the total potential number of spin matches between the spin system and the hypothetical matching residue, the resulting $score_{abs}(ss_x, res_y)$ terms in the above equations also range from 0 to 1. These scores represent an absolute criterion, which is closer to 1 for good spin system to sequence matches and nearer to 0 for poor matches. However, since each spin system must logically correspond to some residue within the protein sequence, only relative sequence fitness is of interest to AutoLink. Thus for each spin system, the scores to each position of the protein sequence are linearly scaled such that each spin system scores 1.0 against at least one position of the protein sequence, if possible. These "relative" sequence matching scores are what is referred to by $score_{rel}(ss_x, res_y)$ in the above equations and are therefore what is used to calculate the overall fit of any given spin system fragment to positions within the protein sequence. Relative sequence matching allows AutoLink to work well with spin systems that contain unusual chemical shifts. Though such exceptional spin systems are generally considered an asset by human spectroscopists, previous automated assignment strategies were generally hindered by them. This is because they rely on absolute criteria which lead to poor scores for spin systems containing unusual chemical shifts. Relative sequence fitting causes AutoLink to evaluate spin-system-to-residue matching in a way more similar to an expert spectroscopist.

It is important to note that AutoLink makes no exceptions in sequence matching for "artifact" spin systems. Artifacts that cannot be labeled as such by the user prior to running AutoLink must be treated as regular spin systems. As is the case for human assigners, AutoLink must use consistency with the protein sequence and with other spin systems in order to rule out artifactual spin systems as possible assignment candidates.

The "empirically determined average" chemical shifts ($chemShift_{resAvg}$) used for sequence matching can come from any source. Particularly useful is nearest-neighbor-based prediction. For this purpose, AutoLink makes use of the parameters presented by Wang and Jardetzky [22] which can take predicted secondary structure into account. The secondary structure prediction can be done using any program [23–25] and the predictions are easily incorporated into AutoLink's calculations through its graphical interface. Alternatively, AutoLink can be instructed to simply use secondary-structure- and/or nearest-neighbor-independent empirically defined average chemical shift values.

Since the matching scores for the fragments are calculated (as described above) by taking the geometric average of the matching scores of the individual spin systems to their hypothetical residues, they also fall in the range of 0–1, with good matches scoring higher than poor matches. Fragment to sequence matching is additionally governed by a user-defined threshold parameter. For any match of a fragment to the protein sequence where the fitness score is less than the matching threshold, the score is rounded to 0, and the fragment is effectively not considered as a potential match to that segment of the protein sequence. In addition to this requirement, each individual spin system of the fragment must also match the protein sequence with a score above the matching threshold. This means that AutoLink only considers fragment-sequence matches at sequence positions where each residue as well as the whole fragment, scored on relative criteria, fit with the user-defined acceptance limit. This limit is usually set to between 0.2 and 0.4, which corresponds to an extremely poor sequence match, so the threshold is generally used only to rule out obviously impossible matches.

Sequence matching can also take into account data from specific labeling, restricting each spin system to match only a subset of the protein residues.

In order to calculate the base priority prime list, the spin system pair scores from the original priority list are multiplied by the fragment score of the hypothetical spin system fragment that will be formed if the spin system link for the pair in the priority list is accepted. Thus, the base priority prime list contains spin system pair scores that are both a function of the relatedness of the spin systems according to the NMR data, and a function of the fit of the hypothetical fragment that each would create to the protein sequence.

Matching of fragments to the protein sequence can also take two other factors into consideration, putative secondary structure elements (assessed based on the C_α chemical shifts), and fitting of other fragments into the protein sequence.

3.2. Calculation of the relative priority prime list (relative hypothesis prioritization)

Once the base priority prime list has been calculated, the list is further modified by biasing potentials. These potentials do not involve any external criteria, but rather involve only processing of the base priority prime list by logical internal criteria. This forms the relative priority prime list, which is the final score list from which AutoLink's decisions are derived. The combination of these biasing potentials gives rise to AutoLink's logical decision-making process.

3.2.1. Score delta bias

Each spin system-pair score in the list is biased according to the next best hypothesis in the list for one of the individual spin systems involved in the pair according to one of the following equations:

$$score'_{A\to B} = (1 - sdb) \times score_{A\to B} + sdb \times [score_{A\to B} \times (score_{A\to B} - score_{A\to ?})] \quad (9)$$

or

$$score'_{A\to B} = (1 - sdb) \times score_{A\to B} + sdb \times [score_{A\to B} \times (score_{A\to B} - score_{?\to B})], \quad (10)$$

where $score_{A\to B}$ refers to the score of spin system "A" → spin system "B" before score delta biasing, $score_{A\to ?}$ refers to the next best score for spin system "A" (spin system "A" → any spin system other than "B"), $score_{?\to B}$ is the next best score for spin system "B" (any spin system other than "A"→"B"), and *sdb* is a user-defined control parameter ranging from 0 to 1. The choice of which equation is used is based on which score delta is higher, $score_{A\to ?}$ or $score_{?\to B}$. Thus, each base priority prime score is multiplied by a factor that is dependent on the difference between that score and the closest related score for one of its spin systems. As an exception, if the score delta biasing for a particular hypothesis would shift the score above the score of an initially better scoring hypothesis that includes the spin system used to calculate the bias, the score delta bias is limited such that the new score is just below the score of the better hypothesis. The result of score delta biasing is that the relative value of the priority prime scores changes. Scores from spin system link hypotheses with lower scoring alternatives are decreased less than those with better alternatives (see Fig. 4).

Score delta biasing, in effect, simulates a "process of elimination." Since spin system link hypothesis accepted in earlier rounds can affect spin-link hypothesis evaluation in later rounds, it is important to accept the hypotheses with the highest *relative* certainty first, reducing the need for re-evaluation later. Hypotheses that initially score well, but have at least one reasonable alternative are disfavored compared to other spin hypothesis for which there is no good alternative. Thus, the hypotheses with few alternatives tend to be accepted in earlier rounds, and are then used in later rounds to aid in evaluating other link hypotheses for which there were initially more than one reasonable possibility.

3.2.2. Repeat bias

During AutoLink rounds, it is possible for the program to get into a cycle, where the acceptance of a set of link hypothesis leads to a re-evaluation of the priority prime lists in such a way as to make the link hypothesis in the set now unfavorable. Thus they can be replaced in a subsequent round. Sometimes, however, the replaced links cause the priority prime scores to once again favor the original links. Unless there is some change in the scoring of the priority prime lists from one cycle to the next, it is possible for AutoLink to get caught in an unending loop, repeatedly trying to sort out a set of links which cannot be properly evaluated until other as of yet unaccepted links are taken into consideration. To solve this problem, AutoLink remembers how many times a particular spin system has been linked to another spin system and introduces a small, user-defined, bias into all of that spin system's priority prime scores. This bias is defined by:

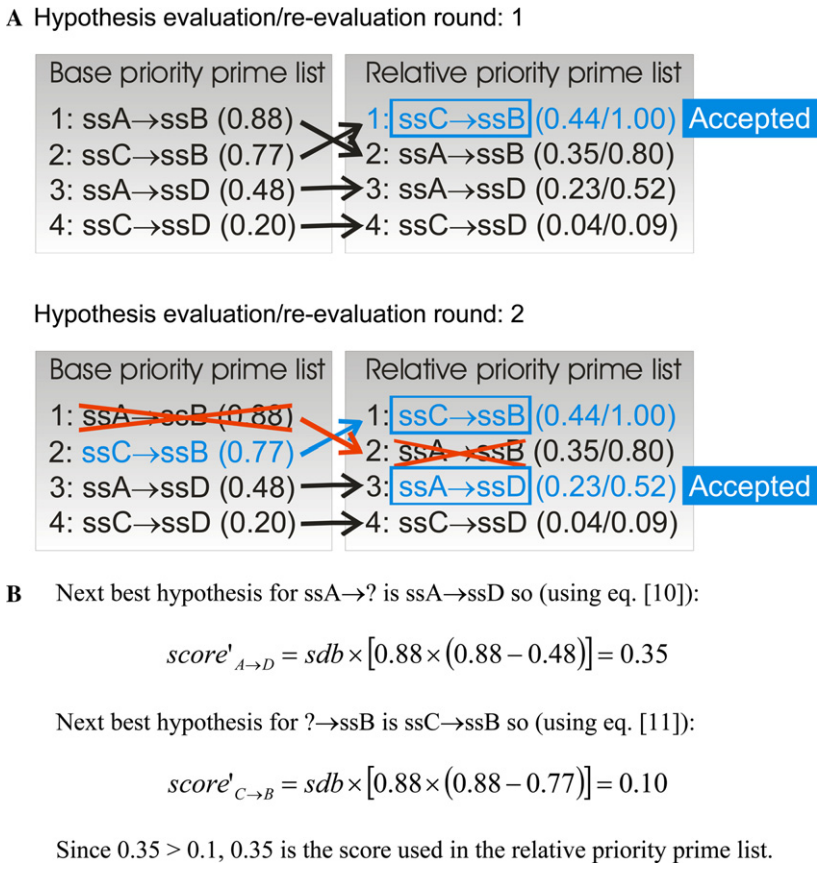
$$score' = score \times rb^{\#ofRepeats}, \quad (11)$$

where *rb* is the user-defined input parameter between 0 and 1, and #*ofRepeats* is the number of times the spin system's link partner was determined or re-determined. #*ofRepeats* is generally set to a value close to but not equal to 1, so that only multiple repeat events have a significant effect. The result of repeat biasing is that link hypotheses that are constantly being re-evaluated are reduced in priority until they score lower than other hypotheses which are not being repeatedly tested. In effect, uncertain hypotheses are "saved for later" until other hypotheses have been evaluated that might be able to be used to sort out the ambiguity. Thus, with a repeat bias setting <1, it is impossible for AutoLink to get stuck in a loop. It simply lowers the priority of the involved hypothesis for later re-evaluation once more of the rest of the hypotheses have been considered.

The combination of score delta biasing and repeat biasing also give AutoLink an unusual ability to "determine what can be determined" in an NMR assignment problem. Score delta biasing causes link hypotheses with more than one reasonable alternative to be disfavored by the program until all of the alternatives can be ruled out. Eventually, once all of the reasonably determinable link hypotheses have been accepted, if any link hypotheses remain, they are accepted and rejected repeatedly, as AutoLink considers each alternative, until the repeat bias increases to a point where the program aborts any further consideration of any of the hypotheses involved (see Section 3.2.4). These remaining hypotheses are thus considered by AutoLink to be undeterminable, since none of the alternatives can be unambiguously chosen.

3.2.3. Random factor bias

All of the priority prime scores can be adjusted according to user-controlled random factors. The random factor generation is controlled by the following equations:



B Next best hypothesis for ssA→? is ssA→ssD so (using eq. [10]):

$$score'_{A\to D} = sdb \times \left[0.88 \times \left(0.88 - 0.48\right)\right] = 0.35$$

Next best hypothesis for ?→ssB is ssC→ssB so (using eq. [11]):

$$score'_{C\to B} = sdb \times \left[0.88 \times \left(0.88 - 0.77\right)\right] = 0.10$$

Since 0.35 > 0.1, 0.35 is the score used in the relative priority prime list.

Fig. 4. (A) Demonstration of RHP principal. For this demonstration, the score delta biasing control parameter (*sdb* in Eqs. (10) and (11)) is assumed to be set to 1. Each gray box represents a priority list both before (left) and after (right) score delta biasing. For simplicity's sake, a model priority list containing only four hypotheses is shown. For typical NMR assignment problems, the length of the priority lists ranges from hundreds to tens of thousands of hypotheses. The link hypotheses in each list are denoted by spin system pairs (ex: ssA → ssB) followed by their priority (fitness) score in parentheses. In the base priority prime lists, the scores shown have already been scaled to 0–1. In the relative priority prime two scores are shown in order to aid the reader in following the calculations. The first of these is the priority score just after relativity biasing. The second score is the relativity biased score re-scaled to values between 0 and 1. Each list is ordered from highest priority score to lowest. Accepted hypotheses are shown in bold. Two rounds of hypothesis evaluation are shown in order to demonstrate the effect of preceding rounds on later rounds. It should be noted that ordinarily more than one hypothesis would be accepted per round by AutoLink and an "unlinking" round would separate the two linking rounds (see Section 3 in the main text). For clarity the process has been simplified by reducing the number of accepted hypotheses per round to one and therefore no unlinking round is possible. (Top) Initially ssA → ssB has the highest priority before score delta biasing (left panel). However, since there are mutually exclusive alternatives (ssC → ssB, ssA → ssD) the priority of the ssA → ssB link hypothesis is reduced (top right panel). The priority of ssC → ssB, on the other hand increases, since the next alternatives for ssC → ? is relatively low priority (0.2). This allows the priority of ssC → ssB to exceed that of both ssA → ssB, allowing it to be accepted preferentially. (Bottom) The acceptance of ssC → ssB in the previous round rules out the ssA → ssB hypothesis (shown with a red cross over it) since ssC → ssB and ssA → ssB are obviously mutually exclusive. Thus, since ssC → ssB was accepted in the first round, AutoLink will now ignore the ssA → ssB hypothesis and is free to choose the ssA → ssD alternative. (B) Sample calculation of relativity biasing for the link hypothesis ssA → ssB from the priority list in Fig. 4A.

$$\begin{aligned} randomFactor = {} & a \times centerRandomFactor + b \\ & \times (posRandomFactor \\ & + negRandomFactor), \end{aligned} \tag{12}$$

$$centerRandomFactor = \frac{rand(nc, -1, 1)}{nc}, \tag{13}$$

$$posRandomFactor = \frac{rand(noc, 0, 1)}{noc \times 2}, \tag{14}$$

$$negRandomFactor = \frac{rand(noc, -1, 0)}{noc \times 2}, \tag{15}$$

where $rand(x, y, z)$ refers to the sum of x pseudo-random numbers generated by the standard C++ random number generator and ranging from y to z, nc is a user-defined parameter indicating the number of "on center" random numbers that will be included in the calculation, noc refers to the number of "off-center" random numbers to be included. a and b are user-defined weighting factors controlling the weighted addition of the off-center and on-center random numbers. With this set of parameters it is possible to design a wide variety of random number distributions. The priority prime score for each residue pair is then modified according to the formula:

$$score' = (1 - rfb) \times score + rfb \times score \times randomFactor^{exp} \times randAmp, \quad (16)$$

where *rfb* is a user-defined weighting factor ranging from 0 to 1 which controls what percentage of the base priority prime score is multiplied by the random factor, *exp* is a user-defined exponent, and *randAmp* is a multiplier that controls the magnitude of the random factor's effect. Inclusion of a significant random component to the priority prime scores causes AutoLink to temporarily re-organize the priority prime list, allowing the program to test marginal hypotheses at random relatively early in its analysis. Since the random factors are re-calculated during every round of linking/unlinking, a hypothesis that was accepted in an early round due to a random increase in its score will typically only be maintained in later rounds if new spin system links are subsequently formed that stabilize the link. Random factor bias can be modeled in terms of human logic in terms of "what if?" type of thinking. Hypotheses can be temporarily accepted in order to investigate either subsequent confirmation or disqualification of the hypotheses. It is noteworthy that inclusion of a significant random component into the priority prime scores will lead to at least some increase in the range of possible results given any particular set of data. Thus, a high degree of random factor biasing is only useful in the context of multiple runs of the AutoLink program, where the comparison of the final results by the user is used to rule out unlikely answers.

3.2.4. Evaluation of the relative priority prime list

Once the relative priority prime list is established, AutoLink begins linking related spin systems. The general approach is that the relative priority prime list is scanned in order of highest to lowest scores and the new link hypotheses with the highest scores (highest priority) are accepted and linked. Since the scores have already been processed by both score delta biasing and repeat biasing, only link hypotheses with no good alternatives are at the top of the list. Thus AutoLink always accepts only hypotheses with *relatively* high certainty. Since the acceptance of link hypotheses affects spin system to protein sequence matching, score delta biasing, and repeat biasing, only a small number of hypotheses should be accepted before the base priority prime list and relative priority prime list must be re-calculated.

Because of this, the linking process is divided into "rounds," with a user-defined maximum number of new links formed each round. Each round, AutoLink evaluates the spin system link hypotheses by calculating the base priority prime list and then the relative priority prime list, and subsequently accepts a user-defined number of hypotheses at the top of the list.

There is one additional criteria required for evaluation of the relative priority prime list—no link hypothesis is accepted if it would preclude another link hypothesis that ranks higher in the list. Since the relative priority prime list is evaluated sequentially, this causes the program to proceed in a manner of another "process of elimination." Spin system link hypotheses can be accepted even if one or both of the spin systems of the pair is involved in a higher scoring pair as long as the higher scoring pair or pairs is itself involved in a yet better scoring pair.

Since the formation of new spin system links affects the relative scoring of previously accepted link hypothesis, each round of linking is followed by a round of unlinking. That is, after each linking round where new spin system link hypotheses are considered, the priority prime lists are re-calculated and, subsequently, the worst scoring links are removed. Thus, the program re-considers its prior conclusions "in the light of new information." For example, if initially spin system "*A*" → spin system "*B*" is linked, and later the spin system "*B*" → spin system "*C*" is accepted (forming the fragment "$A \rightarrow B \rightarrow C$"), then it may be that the link between "*A*" and "*B*" becomes less favorable because the fragment "$A \rightarrow B \rightarrow C$" does not map into the same position in the sequence of the protein as the fragment "$A \rightarrow B$" did. Thus, the link between "*A*" and "*B*" may be broken, especially if spin system "*A*" has other reasonable alternative link hypotheses.

Just as the number of new links per round is controlled by a user-defined parameter, so also is the number of possible "unlinks" per round. The number of links broken per round, however, is restricted such that at least one net positive link will be gained per round whenever possible.

Linking/unlinking rounds proceed alternatingly until either a user-defined number of rounds is reached, or all of the remaining spin system link hypotheses not accepted are below a user-defined threshold. This threshold is generally set to a value far below the range of reasonable spin system links, relying on score delta biasing and repeat biasing to reduce all inconclusive hypotheses below it. AutoLink, therefore, performs spin system linking/unlinking in a directed manner, starting from the best hypothesis and working its way down the relative priority prime list, evaluating and re-evaluating its results until the best net ensemble of links is formed according to the scoring of the relative priority prime list. Since the relative priority prime list takes into account scores from one or multiple spectra as well as fit into the protein sequence, the links formed by AutoLink produce spin system fragments that are consistent with both the NMR data and with the known protein sequence.

3.2.5. Spin system fragment assignment

AutoLink can be directed to assign fragments whose position can be unambiguously mapped into the protein

sequence. The algorithm AutoLink uses to assign fragments to positions within the protein sequence is itself a cycle. At the beginning of the cycle, all of the fragments created by previous linking/unlinking cycles are considered unassigned. Each fragment is then scored (Eq. (7)) against the protein sequence and the fragments are prioritized according to the *n*th root of that score, where *n* is the length of the fragment. The fragment that has the highest priority and that has only one available fitting sequence position according to the sequence fitting threshold (also described in Section 3.1) is assigned first. The *n*th-root-based modification of the fragment score causes AutoLink to bias in favor of forming longer fragments (which are generally more likely to be correct), but still allows a relatively well-fitting fragment to out-compete a longer fragment if its sequence match is significantly better. After the first fragment has been assigned, each fragment is re-scored against the remaining unassigned parts of the protein sequence, and, again, the highest priority fragment that matches only one position within the protein sequence is assigned. This cycle continues until no fragments remain that can be unambiguously assigned to the protein sequence. Fragment assignment is governed by another parameter—no fragment is assigned that is shorter than a user-defined length criterion.

There are two points during the hypothesis evaluation/re-evaluation cycles at which the program can be directed to assign the fragments—at the beginning of each linking round and/or at the beginning of each unlinking round. Assignment of fragments during the linking/unlinking cycles allows AutoLink to take the current fragment assignments into consideration during spin-link hypothesis evaluation. Spin links that lead to fragments whose only favorable matching positions within the protein sequence are assigned to another fragment will be excluded from consideration unless the fragment matches a sequence better than the fragment that is already assigned to that sequence. During the linking phase of the linking/unlinking cycle, this causes AutoLink to never build a fragment which does not have a corresponding sequence fit within the protein sequence that is either unassigned at the time the fragment is created, or assigned to less favorable fragment choice than the new fragment. During the unlinking phase, AutoLink will break any fragment that does not have such an available sequence match at its weakest link. Thus, if new links create a new fragment that matches a part of the protein sequence better than any of the previously existing fragments, any fragments which conflict with the new fragment are broken at their weakest link, and the resulting shorter fragments are freed for re-consideration in subsequent linking/unlinking cycles.

Additionally, AutoLink can be directed to assign fragments only after all of the linking/unlinking cycles have been completed, allowing the user the freedom to resolve conflicts manually.

3.3. Program input/output

AutoLink has a sophisticated graphical user interface (GUI) (Fig. 5), using the powerful library of CARA, which aids the user in both setting up the input controls and interpreting the output spin system fragments and assignment. The input controls are associated with color-changing indicators in order to make it easier for the user to scan the control panels and assess the program's current state. Likewise, the output displays report a color-coded summary that allows the user to assess the quality of the assignments, whether they were made by AutoLink or by the user.

AutoLink can also be used interactively, giving the user more control over the assignment process. It can accept prior assignments from the user and use them to aid in further assignments. It can also accept prior link hypotheses defined by the user and consider them as either absolute truths or suggestions, depending on the user's specification. Interactive use of the program is very helpful in debugging the input spin systems, as mistakes from the human user are quite frequent.

3.3.1. Applications

Test data were downloaded from the Biomagnetic Resonance Bank (BMRB) site at: http://www.bmrb.wisc.edu. The data were obtained from BMRB Accession Nos. 4678 [26], 5106 [27], 5166 [27], 5329 [28], 5335 [29], 5589 [30], 5656 [31], 5691 [32], 5842 [33], 5844 [34], 5845 [35], 5859 [36], 6011 [37], 6052 [38], 6128 [39], 6138 [40], 6176 [41], 6209 [42], 6318 [43], 6341 [44], and 6344 [45]. Of these sets, the peak information was available from 6128, 6176, and 6318. For the remaining 18 test data sets, the published chemical shifts were used to create predicted peaks. For each test, the peak data obtainable from an HNCACB [18–20], a CBCA(CO)NH [21], an HNCO [18–20], an HN(CA)CO [18–20], and a ^{15}N-correlated NOESY [18–20] were simulated. For the simulated peaks in the carbon-correlated spectra, the peak positions were randomized by adding a pseudo-random number between 0 and 0.4 ppm to the carbon chemical shift of each inter-residue crosspeak. This variance was generally consistent with the real peak data for C_αs and C_βs of the data sets 6128, 6176, and 6318, but somewhat above what was observed for the carbonyl crosspeaks. In our own data for ACF RRMs 2 and 3, 0.4 ppm corresponds roughly to the average linewidth of the HNCA crosspeaks. The NOESY crosspeaks were simulated by transferring assigned ^{1}H crosspeaks from their native spin system to the adjacent spin systems and randomizing their positions by up to 0.025 ppm, a value also consistent with the ^{15}N-NOESY data obtained for ACF RRMs 2 and 3.

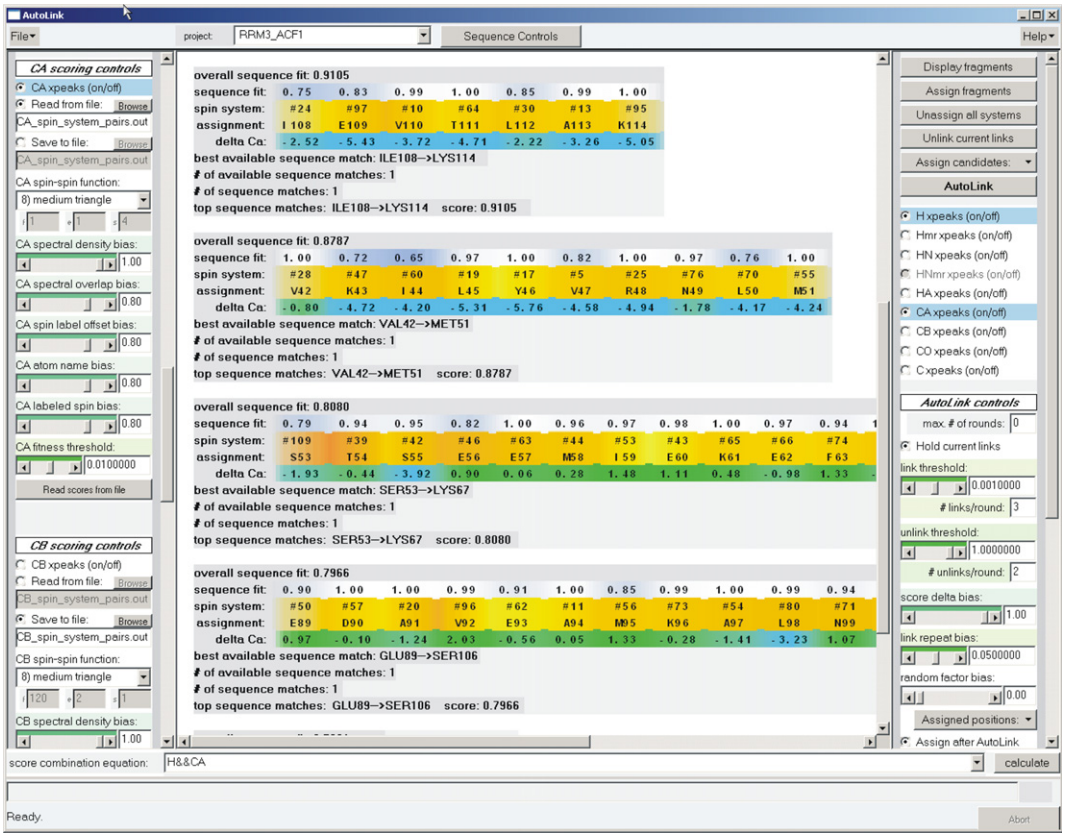


Fig. 5. Main AutoLink program window. The scoring controls for each sub-scoring function are located in the left scroll view. Each sub-score type is controlled by a separate subsection of the window. The C_α scoring controls are currently shown at the top of the window. Each subsection of the window contains a series of file i/o devices as well as potentiometer-type controls for each of the scoring function biases (see Section 2.3.2 in the main text). Color-coded indicators aid the user in quickly scanning the display and avoiding errors. At the bottom of the main display window there is the scoring equation editor. This window contains several selectable equation examples as well as intelligent right-click driven hints to help the user design correct scoring equations. The right scroll window contains the linking controls—that is controls that affect the functions used during the hypothesis evaluation/re-evaluation cycles (see Section 3 in the main text). Thus, these controls handle spin system to sequence matching and the RHP decision-making process. Sequence-specific control functions, such as chemical shift prediction input, are entered by clicking on the "Sequence Control" button at the top of the window and entering the desired settings into a sub-window containing a point-and-click graphical interface. Currently the central view window is showing the "fragment display." This display is primarily used to aid in the assessment of existing spin system links and assignments. Each fragment is represented by three rows of color. The center row shows the spin systems that make up the fragment and their corresponding assignment with the protein sequence if the fragment has been assigned. The center rows are color-coded according to the overall score of each spin system link according to the NMR data. Links that are well supported by the data are yellow while poorer links become more red. Individual sub-scores can be viewed by right clicking between the spin systems. The top row shows the relative fit of each spin system to its position within the protein sequence. The color coding ranges from white (good fit) to blue (marginal fit) to red (poor fit). Right clicking on this row shows the contribution of each spin of the spin system to the overall fit to the protein sequence. The third row of each fragment representation shows the deviation of the C_α observed from the empirically derived mean. The color coding is as follows: blue signifies more β-sheet-like C_αs, green signifies more α-helical C_αs, and intermediate blue-green color signifies random-coil-like C_αs. Additionally, each fragment is listed with its best overall fit to the protein sequence, how many positions within the protein sequence the fragment matches, and what alternative positions are reasonable.

Since peak data were available for test sets 6128, 6176, and 6318, these peaks were used for these tests rather than simulated crosspeaks. A few obvious corrections (such as correction of atomic assignments) were made where necessary. For test sets 6176 and 6318, the HNCA, HNCACB, HNCA, HN(CA)CO, and ^{15}N-NOESY peaks were used since these data were sufficient to obtain 100% of the maximum possible result. For test set 6128 the CCONH [19,20] crosspeaks were additionally used. This allowed the assignments to reach 99% (only 90% assignments were achieved without the CCONH crosspeaks). This data set, however, contains 13 extra spin systems that were somewhat redundant with the assignable spin systems, as well as multiple missing carbonyl and C_β crosspeaks and sparse NOE crosspeaks.

All of the calculations on the test data sets were performed on a Mandax workstation with a 1.21 GHz Athlon processor and 512 MB RAM. These test data sets are available for download from the AutoLink website: http://www.autolink.nmr-software.org/index.htm.

Calculations on experimental data for the RRMs of ACF were performed on a Dell Precision 340 workstation with a 2.53 GHz Pentium 4 processor using all of

the spectra described under Section 2.2. Of the 102 spin systems visible in the spectra of RRM 2, all C_α and $C_{\alpha-1}$ crosspeaks were identifiable and of the $C_{\beta-1}$ crosspeaks three were missing. Of the 91 spin systems that were present in the data for RRM 3, all of the spin systems had identifiable C_α and $C_{\alpha-1}$ crosspeaks, and all but one had an identifiable $C_{\beta-1}$ crosspeak.

It should be noted that while AutoLink has the ability to incorporate secondary structure prediction and nearest-neighbor chemical shift prediction into its analysis, neither of these functions was necessary in order to achieve good results on any of the test or experimental data. Both functions were tested, however, for their effect on the assignments of ACF RRM3.

4. Results

4.1. Automatic assignment of test data from BMRB

Out of all of the test data sets, only one was under-assigned—6128. In no cases were incorrect assignments obtained. Even for data set 6128, which has a high degree of overlap, 99% of the correct assignments was achieved. See Table 3 for a summary of these tests.

4.1.1. Automatic assignment of RRM 2 and 3 from ACF

We applied AutoLink to solve the backbone resonance assignments of RRMs 2 and 3 from ACF. The available NMR spectra for each protein fragment were

Table 3
Summary of results of AutoLink on test and experimental data sets

Molecule (BMRB accession number)	Number of residues	Number of assignable spin systems	Number of assigned spin systems
Simulated data			
RNA polymerase subunit RPB5 (4678)	77	77	77
MTH1743 (5106)	70	70	70
Hemolysin expression modulating protein Hha (5166)	72	72	72
C. elegans protein ZK652.3 (5329)	94	94	94
E. coli protein YacG (5335)	68	64	64
V. cholerae VC0424 (5589)	132	125	125
Staphylococcal protein A (5656)	71	71	71
30S ribosomal protein S28E from *P. horikoshii* (5691)	82	78	78
H. influenza protein IR24 (5842)	134	115	115
S. aureus protein SAV1430 (5844)	91	87	87
S. aureus protein MW2441 (5845)	102	96	96
Antibacterial peptide microcin J25 (5859)	21	21	21
A. thaliana protein At5g22580 (6011)	111	110	110
Haemophilus human protein HR969 (6052)	149	139	139
A. thaliana protein At2g24940.1 (6138)	109	103	103
A. thaliana protein At3g03410.1 (6209)	67	67	67
A. thaliana protein At3g04780.1 (6341)	161	160	160
Human protein HSPCO34 (6344)	143	143	143
Real data—known assignments			
A. thaliana protein At3g01050.1 (6128)	101	114	94
Ubiquitin-like domain of tubulin-folding cofactor B (6176)	120	117	117
A. thaliana thioredoxin h1 (6318)	124	109	109
Real data—unknown assignments			
Human ACF RRM2	108	102	71
Human ACF RRM3	115	91	85

Summary of results of AutoLink on test and experimental data sets. Shown is a list of the protein analyzed with its BMRB accession number followed by the number of residues in the protein construct on which the data were acquired (including vector sequences), the number of assignable spin systems observed in the spectral data, and the number of spin systems assigned by AutoLink. For each of the test sets involving simulated peak data, one spin system was created for every residue assignment published on the BMRB website. Since complete assignments were not available for all of the test proteins, several of the test sets have a maximum number of assignable spin systems that is lower than the number of residues of the protein. In most cases, however, assignments were available for proline residues. Even though these spin systems would not be present in the simulated amide-correlated spectra, they were included in the analysis (if C_α, C_β, or CO resonances were known) to increase the stringency of the tests. In the experimental sets derived from real peak data, however, prolines were unassignable since they did not give rise to amide-correlated crosspeaks. For BMRB test set 6128, although the number of assignable spin systems listed is 114, in fact the actual maximum number of assignments possible was 95, since six of the proteins residues are prolines. AutoLink successfully identified 93 of the assignments exactly as the published assignments in the BMRB data. The program did substitute spin system 108 for spin system 110 as the assignment for the last residue of the protein. This is not regarded as a misassignment, however, as spin systems 108 and 110 are overlapped in the ^{15}N-HSQC and have quite similar ^{13}C chemical shifts. Although the number of assignments for the two RRMs of ACF is rather low, it should be noted that much of the unassigned residues of the proteins are outside of the folded domains. For folded regions of RRM2, and RRM3, not including prolines, 84% of RRM2 was assigned and 97% of RRM3. The main limiting factor in determining the rest of the assignments is missing crosspeaks in the spectra of these molecules.

a ^{15}N-correlated NOESY, an HNCA, a CBCA(CO)NH, and a ^{15}N-HSQC.

For RRM 3, the protein fragment we used was 115 residues long, leading to spectra of reasonable quality and sensitivity with several overlapped spin systems. Spin picking of the spectra was done semi-automatically, with initial peak identification done using standard local maxima assessment by in-house Lua scripts written to work in CARA followed by manual inspection and editing. Of the 115 spin systems, only 90 were visible in the 3D spectra. The program was run using the default control parameters and using the spectral scoring equation: overall score = H && CA. This equation defines a "quasi-AND" operation between each spin system pair's C_α score (obtained from the HNCA peaks) and NOESY (H) score (obtained from the ^{15}N-NOESY peaks) as described in Section 2.3.2.6. AutoLink ran through 120 linking/unlinking cycles with fragment assignment considered in all linking and unlinking rounds. The program terminated itself after approximately 2 h. Each spin system link and assignment was subsequently inspected visually in all three input spectra using CARA. Initial evaluation of the results showed that three spin systems had been picked incorrectly. AutoLink refused to link these spin systems until the corrections were made. Upon correction of the input and re-execution of the program, 79 spin system links were created by AutoLink, all of which appeared probable based on visual inspection. Seventy-five of the linked spin systems were assignable, almost all of which were mapped to the folded domain (assessed based on sequence homology) of the protein. In fact, 94% of the folded domain was assigned, with the remaining unassigned residues being prolines (P30, P68), whose backbone N cannot be assigned from the three input spectra, and/or being from short loops within the RRM domain (L52, R88). Additionally, much of the sequence adjacent to the folded domain was also assigned. Inspection of the remaining unassigned five spin systems showed that they could not confidently be assigned to the missing loop segments and that they were probably part of the leader sequence surrounding and including the HIS(6) tag. Since much of the leader sequence residues did not show up in the spectra, the leader sequence was also unassignable.

For RRM 2, the spectra obtained were of substantially lower quality than those of RRM 3, as the molecule showed limited solubility. To solve this problem, we included a stoichiometric amount of target RNA in the sample. Despite improvement in the solubility, the signal-to-noise remained substantially lower than that for RRM 3, and there was much more overlap evident both in the ^{15}N-HSQC and in the ^{13}C dimensions of the 3D spectra. Initial spin system picking was performed as for RRM 3 above and AutoLink was run on the input using the same input parameters. The output was inspected with AutoLinks's fragment display with CARA's spectrum viewing tools and some obvious errors in the spin system peak identification were corrected. Upon re-running the program ~80% of the non-proline residues of the RRM domain could be assigned. With a few more rounds of inspection, spin system correction and program execution, 84% was assigned. The remaining spin systems could only be assigned with marginal certainty or were unassignable due to poor representation of the spin systems in the spectra. Overall, the assignment of RRM 2 took about 2 days time with most of the time being occupied by the computer. The assigned residues were C29-I36, K38-V61, S64-D67, G73-K91, and R96-W108.

The overall results obtained from both RRMs 1 and 2 verify AutoLink's ability to discriminate between determinable and undeterminable parts of the assignment problem. As a further test of AutoLink's "solvability" discrimination, it was also run on RRM 3 using only two of the three available 3D spectra. The results varied according to which spectra were included, but were consistent in that, in each case, substantially less complete assignments were obtained. This is expected since the three spectra combined are considered to be the minimum required amount of NMR data to obtain objective backbone resonance assignments of most proteins of this size. For those assignments that were obtainable with fewer spectra, however, the results generally were consistent with those obtained when all three spectra were included. The best case was when the NOESY data were excluded from the analysis (leaving only an HNCA and a CBCA(CO)NH) but when nearest neighbor-predicted chemical shifts were used based on secondary structure predictions from YASPIN [22]. Only the secondary structure predictions for sequences with the highest confidence ratings were used. Nevertheless, AutoLink was still able to assign 73 spin systems correctly with no misassignments. This is good evidence that AutoLink's solvability discrimination causes the program to only report relatively certain results.

5. Discussion

5.1. Overall approach to using AutoLink

The usual paradigm to proper use of AutoLink is to first start with stringent input parameters for the first run, inspect the results, and then run subsequent analysis with either less stringent parameters or more stringent sequence restrictions. Subsequent runs of AutoLink are generally used to try and get the program to compensate for human errors in the input spin systems. Thus, at first AutoLink is run with all of the biases set to 1, except for the repeat bias (which is always set to around 0.95) and the spin system label biases for the

NOESY peaks (which are set to 0 since the NOESY peaks are not labeled prior to backbone resonance assignment). The carbon matching functions are generally set to medium-width triangular spin–spin comparison functions, while the H–H (NOESY) scoring function is best set to a narrow parabola (this is because the NOESY spectra are comparatively high-peak-density). Furthermore, the first run of AutoLink is usually executed with a stringent scoring equation, which is either a multiplication of or a quasi-AND of the input spectrum score types. Later runs, if necessary, generally are less stringent, but benefit from maintaining the results of the prior stringent runs by instructing to program to not change the spin system links and assignments from previous runs. A wide variety of possibilities exists for the relaxation of requirements for subsequent runs, but the most useful are substitution of ANDs with ORs or AND/ORs and multiplication with addition in the scoring equation. This allows AutoLink to compensate for either wrong data or a lack of data in one spectrum if there is a clear match in another spectrum. Alternatively the user can adjust the biasing controls of the individual scoring functions to allow for the possibility of human error. Though the initial run of AutoLink takes a few hours, the subsequent runs are usually done in a few minutes. This is because most of the spin system links are pre-formed by the earlier AutoLink runs, so the program has much less to consider in subsequent runs.

It is, in fact, also possible to direct AutoLink to effectively combine stringent and non-stringent score requirements in a single run of the program simply by defining a more complex scoring equation with higher coefficients on more stringent scoring terms and lower coefficients on less stringent scoring terms. The biasing potentials can also be used to contribute to this ''one-run'' approach by setting them to values >0.5 but <1.0. This causes the program to bias heavily in favor of the user's interpretation of the input data while still allowing at least some room for error. In practice, however, such complicated approaches and often even subsequent runs of AutoLink are not necessary, as the program usually gives enough results to allow the user to correct input mistakes and finish the assigning semi-automatically quickly using the graphical output display and CARA.

6. Conclusion

AutoLink demonstrates an unusual level of intelligence with regard to obtaining backbone resonance assignments from NMR data. Its ability to discriminate between solvable and unsolvable parts of assignment problems gives it the ability to produce high-confidence solutions. Though AutoLink's logical capabilities are advanced, further improvement is possible. Currently AutoLink must work from user-identified spin systems and spins. We performed the initial identification of these data elements with an automatic peak picker using a standard local maxima approach. Such a device, however, invariably produces mistakes, and therefore the spin systems must be visually inspected by the user prior to using AutoLink.

Since AutoLink uses scoring functions in order to evaluate the relatedness of spins and spin systems, it is possible to add additional scoring functions that perform correlations directly on the spectral data itself. This would effectively allow AutoLink to be able to ''see'' the data, reducing the need for accurate peak identification prior to running the program. This type of spectral evaluation combined with the integration of an RHP-based automatic peak picking and spin system identification algorithm could allow the program to work completely from raw data, possibly with no prior spin system input from the user at all.

It should also be possible to develop an RHP-based strategy to assign side chains.

AutoLink is available for free download at http://www.autolink.nmr-software.org/index.htm with pre-optimized default parameters preset in the program. A basic user manual is also available at http://www.autolink.nmr-software.org/index.htm/instructions.

Acknowledgments

The authors extend special thanks to Dr. Christophe Maris for providing the samples of ACF RRMs 2 and 3. Salary for J.E.M. was provided by Professor Frederic Allain and the Swiss Federal Institute of Technology (ETH).

References

[1] J. Liu, H. Hegyi, T.B. Acton, G.T. Montelione, B. Rost, Automatic target selection for structural genomics on eukaryotes, Proteins 56 (2004) 188–200.
[2] P. Deloukas et al., The DNA sequence and comparative analysis of human chromosome 20, Nature 414 (2001) 865–871.
[3] S. Kim, T. Szyperski, GFT NMR, a new approach to rapidly obtain precise high-dimensional NMR spectral information, J. Am. Chem. Soc. 125 (2003) 1385–1393.
[4] J.P. Linge, M. Habeck, W. Rieping, M. Nilges, ARIA: Automated NOE assignment and NMR structure calculation, Bioinformatics 19 (2003) 315–316.
[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] H.N. Moseley, D. Monleon, G.T. Montelione, Automatic determination of protein backbone resonance assignments from triple

resonance nuclear magnetic resonance data, Methods Enzymol. 339 (2001) 91–108.
[8] S.G. Hyberts, G. Wagner, IBIS—A tool for automated sequential assignment of protein spectra from triple resonance experiments, J. Biomol. NMR 26 (2003) 335–344.
[9] B.E. Coggins, P. Zhou, PACES: Protein sequential assignment by computer-assisted exhaustive search, J. Biomol. NMR 26 (2003) 93–111.
[10] R. Keller, K. Wuthrich, Computer-aided resonance assignment (CARA). Available from: <http://www.nmr.ch>.
[11] L.A. Zadeh, Fuzzy sets, Inform. Control 8 (1965) 338–353.
[12] L.A. Zadeh, K.-S. Fu, K. Tanaka, M. Shimura (Eds.), Fuzzy Sets and Their Applications to Cognitive and Decision Processes, Academic Press, New York, 1975.
[13] D.H. Rouvray (Ed.), Fuzzy Logic in Chemistry, Academic Press, New York, 1997.
[14] D. Dubois, H. Prade, Fuzzy Sets and Systems: Theory and Applications, Academic Press, New York, 1980.
[15] F.M. McNeill, E. Thro, Fuzzy Logic: A Practical Approach, Academic Press, New York, 1994.
[16] E. Cox, The Fuzzy Systems Handbook: A Practitioner's Guide to Building, Using, and Maintaining Fuzzy Systems, Academic Press, New York, 1994.
[17] A. Mehta, M.T. Kipler, N.E. Sherman, D.M. Driscoll, Molecular cloning of apobec-1 complementation factor, a novel RNA binding protein involved in the editing of apolipoprotein B mRNA, Mol. Cell. Biol. 20 (2000) 1846–1854.
[18] K. Wüthrich, NMR of Proteins and Nucleic Acids, Wiley, New York, NY, 1986.
[19] G.M. Clore, A.M. Gronenborn (Eds.), NMR of Proteins, CRC Press, Boca Raton, FL, 1993.
[20] J. Cavanagh, W.J. Fairbrother, A.G. Palmer III, N.J. Skelton, Protein NMR Spectroscopy: Principles and Practice, Academic Press, San Diego, 1996.
[21] S. Grzesiek, A. Bax, Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR, J. Am. Chem. Soc. 114 (1992) 6291–6293.
[22] Y. Wang, O. Jardetzky, Investigation of the neighboring residue effects on protein chemical shifts, J. Am. Chem. Soc. 124 (2002) 14075–14084.
[23] K. Lin, V.A. Simossis, W.R. Taylor, J. Heringa, A simple and fast secondary structure prediction method using hidden neural networks, Bioinformatics 21 (2005) 152–159.
[24] K.P. Wu, H.N. Lin, J.M. Chan, T.Y. Sun, W.L. Hsu, A hybrid protein secondary structure prediction algorithm—a knowledge-based approach, Nucleic Acids Res. 32 (2004) 5059–5065.
[25] L.J. McGuffin, K. Bryson, D.T. Jones, The PSIPRED protein structure prediction server, Bioinformatics 16 (2000) 404–405.
[26] A. Yee, V. Booth, A. Dharamsi, A. Engel, A.M. Edwards, C.H. Arrowsmith, Solution structure of the RNA polymerase subunit RPB5 from *Methanobacterium thermoautotrophicum*, Proc. Natl. Acad. Sci. USA 97 (2000) 6311–6315.
[27] A. Yee, X. Chang, A. Pineda-Lucena, B. Wu, A. Semesi, B. Le, T. Ramelot, G.M. Lee, S. Bhattacharyya, P. Gutierrez, A. Denisov, C.H. Lee, J.R. Cort, G. Kozlov, J. Liao, G. Finak, L. Chen, D. Wishart, W. Lee, L.P. McIntosh, K. Gehring, M.A. Kennedy, A.M. Edwards, C.H. Arrowsmith, An NMR approach to structural proteomics, Proc. Natl. Acad. Sci. USA 99 (2002) 1825–1830.
[28] J.R. Cort, Y. Chiang, D. Zheng, G.T. Montelione, M.A. Kennedy, NMR structure of conserved eukaryotic protein ZK652.3 from *C. elegans*: a ubiquitin-like fold, Proteins 48 (2002) 733–736.
[29] T.A. Ramelot, J.R. Cort, A.A. Yee, A. Semesi, A.M. Edwards, C.H. Arrowsmith, M.A. Kennedy, NMR structure of the *Escherichia coli* protein YacG: a novel sequence motif in the zinc-finger family of proteins, Proteins 49 (2002) 289–293.
[30] T.A. Ramelot, S. Ni, S. Goldsmith-Fischman, J.R. Cort, B. Honig, M.A. Kennedy, Solution structure of *Vibrio cholerae* protein VC0424: a variation of the ferredoxin-like fold, Protein Sci. 12 (2003) 1556–1561.
[31] D. Zheng, Y.J. Huang, H.N. Moseley, R. Xiao, J. Aramini, G.V. Swapna, G.T. Montelione, Automated protein fold determination using a minimal NMR constraint strategy, Protein Sci. 12 (2003) 1232–1246.
[32] J.M. Aramini, Y.J. Huang, J.R. Cort, S. Goldsmith-Fischman, R. Xiao, L.Y. Shih, C.K. Ho, J. Liu, B. Rost, B. Honig, M.A. Kennedy, T.B. Acton, G.T. Montelione, Solution NMR structure of the 30S ribosomal protein S28E from *Pyrococcus horikoshii*, Protein Sci. 12 (2003) 2823–2830.
[33] T.A. Ramelot, J.R. Cort, S. Goldsmith-Fischman, G.J. Kornhaber, R. Xiao, R. Shastry, T.B. Acton, B. Honig, G.T. Montelione, M.A. Kennedy, Solution NMR structure of the iron-sulfur cluster assembly protein U (IscU) with Zinc bound at the active site, J. Mol. Biol. 344 (2004) 567–583.
[34] M.C. Baran, J.M. Aramini, Y.J. Huang, R. Xiao, T.B. Acton, L.-Y. Shih, G.T. Montelione, Solution structure determination of the *Staphylococcus aureus* hypothetical protein SAV1430, BMRB website: http://www.bmrb.wisc.edu (in preparation).
[35] J.R. Cort, G.T. Montelione, M.A. Kennedy, Solution structure of *S. aureus* protein MW2441, BMRB website: http://www.bmrb.wisc.edu.
[36] M.J. Bayro, J. Mukhopadhyay, G.V. Swapna, Y.J. Huang, L.C. Ma, E. Sineva, P.E. Dawson, G.T. Montelione, R.H. Ebright, Structure of antibacterial peptide microcin J25: a 21-residue lariat protoknot, J. Am. Chem. Soc. 25 (2003) 12382–12383.
[37] G. Cornilescu, C.C. Cornilescu, Q. Zhao, R.O. Frederick, F.C. Peterson, S. Thao, J.L. Markley, Solution structure of a homodimeric hypothetical protein, At5g22580, a structural genomics target from *Arabidopsis thaliana*, J. Biomol. NMR 29 (2004) 387–390.
[38] T.A. Ramelot, G.T. Montelione, M.A. Kennedy, Backbone and side chain ^{1}H, ^{13}C, and ^{15}N chemical shift assignments for *Haemophilus* human protein HR969, BMRB website: http://www.bmrb.wisc.edu (in preparation).
[39] D.A. Vinarov, B.L. Lytle, F.C. Peterson, E. Tyler, B.F. Volkman, J.L. Markley, Eukaryotic cell-free structural genomics: NMR structure of a beta-grasp fold protein from *Arabidopsis thaliana*, BMRB website: http://www.bmrb.wisc.edu (in preparation).
[40] J. Song, D.A. Vinarov, E.M. Tyler, M.N. Shahan, R.C. Tyler, J.L. Markley, Hypothetical protein At2g24940.1 from *Arabidopsis thaliana* has a cytochrome b5 like fold, J. Biomol. NMR 30 (2004) 215–218.
[41] B.L. Lytle, F.C. Peterson, S.H. Qiu, M. Luo, Q. Zhao, J.L. Markley, B.F. Volkman, Solution structure of a ubiquitin-like domain of tubulin-folding cofactor B, J. Biol. Chem. 279 (2004) 46787–46793.
[42] J. Song, Q. Zhao, S. Thao, R.O. Frederick, J.L. Markley, Solution structure of a calmodulin-like calcium-binding domain from *Arabidopsis thaliana*, J. Biomol. NMR 30 (2004) 451–456.
[43] F.C. Peterson, B.L. Lytle, S. Sampath, D.A. Vinarov, E.M. Tyler, M. Shahan, J.L. Markley, B.F. Volkman, Solution structure of thioredoxin h1 from *Arabidopsis thaliana*, BMRB website: http://www.bmrb.wisc.edu (in preparation).
[44] J. Song, R.C. Tyler, M.S. Lee, J.L. Markley, Solution structure of At3g04780.1, an *Arabidopsis* ortholog of the C-terminal domain of human thioredoxin-like protein, BMRB website: http://www.bmrb.wisc.edu (in preparation).
[45] T.A. Ramelot, M.A. Kennedy, Solution NMR structure of human protein HSPCO34. Northeast Structural Genomics Target HR1958, BMRB website: http://www.bmrb.wisc.edu.