Available online at www.sciencedirect.com
SCIENCE @ DIRECT®

Journal of Magnetic Resonance 170 (2004) 263–277

JMR Journal of Magnetic Resonance

www.elsevier.com/locate/jmr

A generalized approach to automated NMR peak list editing: application to reduced dimensionality triple resonance spectra

Hunter N.B. Moseley[a,d], Nadeem Riaz[a,d], James M. Aramini[a,d], Thomas Szyperski[b,d], Gaetano T. Montelione[a,c,d,*]

[a] Department of Molecular Biology and Biochemistry, Center for Advanced Biotechnology and Medicine, Rutgers University, Piscataway, NJ 08854, USA
[b] Department of Chemistry, University at Buffalo, The State University of New York Buffalo, NY 14260, USA
[c] Department of Biochemistry and Molecular Biology, Robert Wood Johnson Medical School, Piscataway, NJ 08854, USA
[d] The Northeast Structural Genomics Consortium, USA

Received 7 February 2004; revised 2 June 2004
Available online 30 July 2004

Abstract

We present an algorithm and program called *Pattern Picker* that performs editing of raw peak lists derived from multidimensional NMR experiments with characteristic peak patterns. *Pattern Picker* detects groups of correlated peaks within peak lists from reduced dimensionality triple resonance (RD-TR) NMR spectra, with high fidelity and high yield. With typical quality RD-TR NMR data sets, *Pattern Picker* performs almost as well as human analysis, and is very robust in discriminating real peak sets from noise and other artifacts in unedited peak lists. The program uses a depth-first search algorithm with short-circuiting to efficiently explore a search tree representing every possible combination of peaks forming a group. The *Pattern Picker* program is particularly valuable for creating an automated peak picking/editing process. The *Pattern Picker* algorithm can be applied to a broad range of experiments with distinct peak patterns including RD, G-matrix Fourier transformation (GFT) NMR spectra, and experiments to measure scalar and residual dipolar coupling, thus promoting the use of experiments that are typically harder for a human to analyze. Since the complexity of peak patterns becomes a benefit rather than a drawback, *Pattern Picker* opens new opportunities in NMR experiment design.

Keywords: Automated NMR data analysis; Depth-first search; *Pattern Picker*; Peak list editing; Reduced dimensionality

1. Introduction

Advances in sample preparation, hardware for data collection pulse sequence development and experiment design, and software for automated analysis provide a significant reduction in the time necessary to generate biomolecular NMR structures [11–13,21–23,35]. International efforts in structural genomics centering on "high throughput" analysis of protein resonance assignments and solution NMR protein structure determination are spearheading advances that reduce the time, effort, and expense to generate protein resonance assignments and structures [4,14,21,26]. Spurred in part by the demand and opportunities of structural genomics, many steps in the process of NMR data collection and analysis have been streamlined and greatly improved. Still, the first spectral analysis step in the NMR protein structure determination process, peak picking and peak list edit-

[*] Corresponding author. Fax: 1-732-235-5633.
E-mail addresses: hunter@cabm.rutgers.edu (H.N.B. Moseley), guy@cabm.rutgers.edu (G.T. Montelione).

1090-7807/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmr.2004.06.015

ing, requires an enormous amount of human attention. In the case of protein resonance assignments, peak picking/editing has become the most time consuming and labor intensive step. Moreover, the overall robustness of the assignment and structure determination process is completely reliant on the veracity of the peaks selected in this step.

Peak picking involves distinguishing real peaks from noise and artifacts. Some peaks are not significantly stronger than the noise level and at least a few peaks in most protein spectra are overlapped. In addition, multidimensional NMR spectra often exhibit artifacts of baseline distortions, intense solvent lines, ridges, and/or sinc wiggles. These problems are sometimes exacerbated by different processing methods that can dramatically affect lineshape, intensity, and resolution of peaks as well as the severity of spectral artifacts.

Most automated peak pickers [8–10,12,17,24] rely on properties of an individual peak along with a model of the noise in the spectrum to determine whether a peak is valid or not; though, one approach has looked at comparative properties of doublets [2]. These programs test for features of a peak that are, for the most part, the same that a human analyzer would use: the intensity of the peak in comparison to local noise, and the shape of the peak. Many programs perform limited peak editing by filtering one peak list against another in comparable dimensions (see for example [18]). The better peak pickers often obtain reliable results on 2D spectra; however, results on 3D spectra normally require significant manual intervention.

One popular peak picker, the contour approach to peak picking (CAPP),[1] relies primarily on peak shape [9]. After CAPP generates a contour plot, it calculates ellipses that best fit the contours. CAPP then detects potential ridges before finally testing the ellipsoid model of each potential peak against cut-off conditions. Although the results for 2D spectra are generally quite good, 3D spectra still require manual editing. AUTOPSY is another very successful peak picker [17]. It has methods to deal with overlap and deviations from ideal Lorentzian lineshapes, and also takes advantage of symmetry peaks present in some spectra (e.g., COSY, NOESY). However, the strategies it employs generally do not perform well with spectra of more than two-dimensions due to low digital resolution in the indirect dimensions [17].

Some experiments encode information in groups of peaks and represent a unique challenge to standard peak picking algorithms, since the characteristic of the entire group of peaks is required to extract the encoded information. In particular, reduced dimensionality (RD) triple resonance experiments [5,7,28,30–35] encode N + 1 chemical shifts into three N-dimensional peaks, one central peak and two outer doublet peaks, as shown in Fig. 1. Unambiguous identification of the doublet requires the presence of the central peak [32,33]. Hence, a group of one central and two doublet N-dimensional peaks needs to be identified to obtain all $N + 1$ chemical shifts. Technically, the joint sampling of a chemical shift with another phase-sensitively detected dimension gives rise to a cosine-modulation of the transfer amplitude. This yields the doublet peaks (Fig. 1), representing the *additionally encoded information*. Related G-matrix Fourier transformation (GFT) experiments encode $N + K$ dimensional spectral information into a group of $2^{K+1}-1$ N-dimensional peaks [15,16]. Several benefits are derived from such RD and GFT experiments, including reduced data collection times and richer patterns of peaks, which should be more amenable to automated peak editing algorithms.

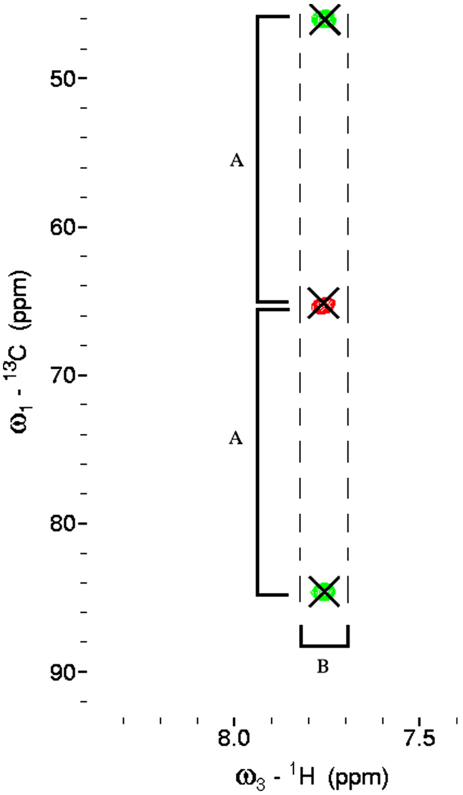


Fig. 1. The RD-TR NMR *Pattern Model*. The center red peak is the central peak of the pattern. The outer two green peaks are the doublet peaks. The two distances marked A show the Doublet Symmetry *Equivalence Relation* of the pattern. The range marked B shows the $^{1}H^{N}$-Matching *Equivalence Relation* of the pattern. An analogous range in the nitrogen dimension (not shown) represents the ^{15}N-Matching *Equivalence Relation*.

[1] *Abbreviations used:* CAPP, contour approach to peak picking; DFS, depth-first search; GFT, G-matrix Fourier transformation; HSQC, heteronuclear single quantum correlation; IPAP, in-phase/anti-phase; RD, reduced dimensionality; RD-TR, reduced dimensionality triple resonance; RDC, residual dipolar coupling; RMS, root mean square.

At first glance, selecting a group of peaks would appear to make peak picking and peak list editing harder. However, the presentation of spectral data in groups of peaks with characteristic relationships between components opens up new avenues for automating the process of peak picking and editing. The constituent peaks in RD NMR spectra satisfy a set of relationships depicted in a common pattern. This set of relationships provides additional constraints to verify the veracity of each peak in a group. Instead of selecting one peak at a time, one can select a group of peaks that, together, fulfill the pattern and, hence, mutually support each other.

We have created an algorithm to edit a raw list of peaks by selecting groups of peaks that fit a defined pattern. This algorithm is applicable to any experiment providing a unique pattern of peaks. We have implemented this algorithm in a C++ program called *Pattern Picker* for the x86-Linux platform. *Pattern Picker* takes in raw lists of peaks and detects groups of peaks that satisfy a set of relationships describing some pattern. The program generates both a list of all the "*Peak Groups*" along with their constituent peaks and a list of the decoded information (e.g., the projected dimension of an RD experiment) contained in each *Peak Group* . In addition, it generates a report that displays detailed information on each group of peaks selected, the results of additional tests on each group, and statistics on the sample of *Peak Groups*. The program is designed to be very flexible with respect to the peak patterns it can recognize and includes facilities to easily craft new patterns. The version of *Pattern Picker* described here can recognize patterns characteristic of several RD triple resonance (RD-TR) NMR experiments. The results presented in this paper demonstrate the high accuracy and robustness of *Pattern Picker* in editing peak lists derived from a representative set of RD-TR NMR experiments.

2. Methods

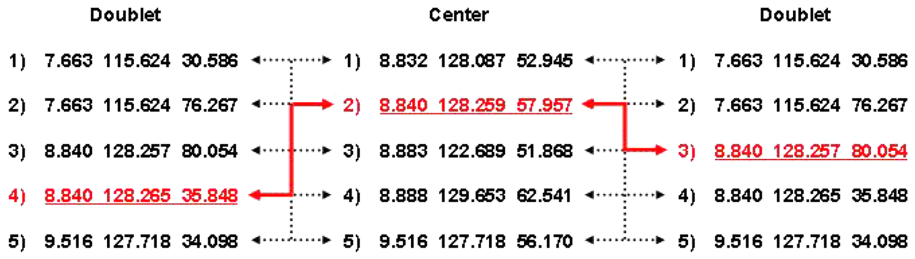
2.1. Algorithm overview

Raw peak lists contain significant numbers of artifactual (systematic errors) and noise peaks (non-systematic errors). Instead of relying only on properties intrinsic to a particular peak such as its intensity or line/peak shape, *Pattern Picker* edits raw peak lists generated by simple peak picking algorithms with the aim of distinguishing "false" peaks from real peaks by selecting those groups of peaks that satisfy a specified set of relationships. These relationships are the core representation of a pattern and provide additional constraints to verify the veracity of each peak in a group. Thus, peaks within a group, satisfying specified pattern relationships, provide mutual support for each other. As the number of these relationships increases, the probability that an invalid *Peak Group* will satisfy all of them decreases, and so the probability of selecting an invalid group, and thus an invalid peak, decreases.

For a particular NMR experiment, *Pattern Picker* creates *Peak Groups* of size *n*, the expected number of peaks in the pattern. For an RD-TR NMR experiment the expected number of peaks or pattern size is three. Other experiments may represent spectral data in patterns of different sizes. Some NMR experiments result in spectra that have patterns without a fixed size. Such patterns would have a variable size within set ranges.

Searching for variable size *Peak Groups* that satisfy a set of relationships easily lends itself to a depth-first search with short-circuiting [27]. A depth-first search (DFS) methodically explores a search tree representing every single possible *Peak Group* and partial *Peak Group*. A DFS starts by selecting the first peak in the group, then the second peak, then the next peak, until it has selected all peaks for one particular group. Restated, the DFS traverses a search tree, where each node in the tree represents a selected peak in a *Peak Group* until it reaches the end node or depth of the tree, which corresponds to the size of the *Peak Group*. If the constructed group satisfies all of the relationships of the experiment, it is added to a *Potential List* of *Peak Groups*. If the last peak added violates a relationship, the algorithm tries another peak for this position in the group (i.e., another sibling node in the search tree). This represents a short-circuiting of the current search path in the tree. If it cannot find a suitable peak, it backtracks (moves up to the parent node) and selects another peak for the previous position in the group (i.e., sibling node of the parent) and then proceeds forward again. *Short-circuiting* avoids exploring unfruitful paths since all peaks selected so far in the search must pass all testable relationships. Paths with earlier failed peak choices are never explored, pruning the tree of exploration. Normally, a DFS requires exponential running time, however, the short-circuiting makes the algorithm tractable for data sets like those generated in RD-TR and other biomolecular NMR experiments. Fig. 2 illustrates the process of finding a *Potential Peak Group* using a DFS with short-circuiting.

After creating the *Potential List* of *Peak Groups*, each group is scored. The score is dependent on the particular pattern used. *Peak Groups* from the *Potential List* are tested for consistency, one at a time, in descending order of their score (best first). The consistency test compares the selected potential *Peak Group* against the *Confirmed List* of *Peak Groups* to make sure that peaks are being appropriately reused. If a *Peak Group* passes this test, it is added to the *Confirmed List*. Once all *Peak Groups* from the *Potential List* are tested, *Pattern Picker* performs a statistical analysis on the *Confirmed List* to determine if the strictness of the *Relationship Tests* should



Group of Peaks Tested	Reason for Acceptance/Rejection
D1,2-C*	Fails ^{1}H^{N}-matching *Equivalence Relation*
D3-C1,2	Fails Peak Ordering *Relationship Test*
D3-C3,4	Fails ^{15}N -matching *Equivalence Relation*
D3-C5	Fails ^{1}H^{N} -matching *Equivalence Relation*
D4-C1-D1,2	Fails ^{1}H^{N} -matching *Equivalence Relation*
D4-C1-D3	Fails Doublet Symmetry *Equivalence Relation*
D4-C1-D5	Fails ^{1}H^{N} -matching *Equivalence Relation*
D4-C2-D1,2	Fails ^{1}H^{N} -matching *Equivalence Relation*
D4-C2-D3	**Passes all *Relationship Tests***
D4-C2-D5	Fails ^{1}H^{N} -matching *Equivalence Relation*
D4-C3,4	Fails ^{15}N -matching *Equivalence Relation*
D4-C5	Fails ^{1}H^{N} -matching *Equivalence Relation*
D5-C1,2,3,4	Fails ^{1}H^{N} -matching *Equivalence Relation*
D5-C5-D*	Fails ^{1}H^{N} -matching *Equivalence Relation*

Fig. 2. A depth first search with short circuiting finding a *Potential Peak Group* using a RD-TR NMR *Pattern Model*. The top of the figure shows a portion of the doublet and center peaks from the HACAcoNH spectrum of *E. coli* yggU. The dotted lines represent all the search paths of the DFS. The solid line and underlined peaks represent the single *Potential Peak Group* that passes all the *Relationship Tests* in the RD-TR NMR *Pattern Model*. The bottom part of the figure shows all the groups of peaks tested during the DFS and the reason for rejection (or acceptance). The partial groups of peaks demonstrate the short circuit testing which prevents the DFS from exploring unfruitful paths.

be adjusted and the process repeated. After several iterations, the statistical tests converge and *Pattern Picker* outputs a list of *Peak Groups*, encoded frequency information, and a report detailing each *Peak Group* and how well it conforms to each *Relationship Test*.

2.2. Pattern Models

Pattern Picker is designed to identify *Peak Groups* from any experiment that provides multidimensional spectral data as a pattern of peaks. Different experiments will have drastically different patterns, thus *Pattern Picker* uses an abstract notion of a pattern called a *Pattern Model*. However, sets of similar experiments often use the same *Pattern Model*. For instance, the RD-TR experiments HACAcoNH (H$^{\alpha}$ *additionally encoded information* from the C$^{\alpha}$ dimension) and HABCABcoNH (H$^{\alpha}$/H$^{\beta}$ *additionally encoded information* from the C$^{\alpha}$/C$^{\beta}$ dimension) [31,33–35] both share the same underlying *Pattern Model*. A *Pattern Model* provides *Pattern Picker* with all the necessary experiment-specific information to select *Peak Groups* from the raw peak lists, create potential *Peak Groups*, check consistency, add *Peak Groups* to a *Confirmed List*, decode encoded information, detect errors and peak overlap, and generate a report. The major components of a *Pattern Model* are outlined in Table 1.

The *Relationship Tests* are the most important part of a *Pattern Model*. They describe the pattern of peaks within a group encoded by a particular NMR experiment (examples of *Relationship Tests* from the RD-TR NMR *Pattern Model* are shown in Table 2). Many,

Table 1
Major components of a *Pattern Model*

Component	Description
Set of *Relationship Tests*	Core description of the expected pattern.
Allowed Missing Method	Determines if an incomplete *Peak Group* is permissible.
Consistency Method	Ensures consistency of the *Confirmed List*.
Decoding Method	Decodes information in a *Peak Group*.
Error Detection Method[a]	Tests *Confirmed List* for possible error *Peak Groups*.
Outlier Detection Method[a]	Tests for outlier *Peak Groups* using looser constraints.

[a] These methods are optional.

Table 2
Relationship Tests and associated *Equivalence Relations* for the RD-TR NMR *Pattern Model*

Test/relation name	Pseudo code description[a]
Equivalence Relationship Tests	
^{15}N-Matching Equivalence Relation	
Center–Doublet1	abs(C.^{15}N.PPM – D1.^{15}N.PPM) < ^{15}N_cutoff
Center–Doublet2	abs(C.^{15}N.PPM – D2.^{15}N.PPM) < ^{15}N_cutoff
Doublet1–Doublet2	abs(D1.^{15}N.PPM – D2.^{15}N.PPM) < ^{15}N_cutoff
^{1}H^{N}-Matching Equivalence Relation	
Center–Doublet1	abs(C.^{1}H^{N}.PPM – D1.^{1}H^{N}.PPM) < ^{1}H^{N}_cutoff
Center–Doublet2	abs(C.^{1}H^{N}.PPM – D2.^{1}H^{N}.PPM) < ^{1}H^{N}_cutoff
Doublet1–Doublet2	abs(D1.^{1}H^{N}.PPM – D2.^{1}H^{N}.PPM) < ^{1}H^{N}_cutoff
Doublet Symmetry Equivalence Relation	S = D1.^{13}C.PPM + D2.^{13}C.PPM – 2*C.^{13}C.PPM; abs(S) < sym_cutoff
Non-Equivalence Relationship Tests	
RMS peak filtering	
Center RMS	0 < C.RMSFit < 500
Doublet1 RMS	0 < D1.RMSFit < 500
Doublet2 RMS	0 < D2.RMSFit < 500
Linewidth peak filtering	
Center ^{13}C linewidth	10 Hz < C.^{13}C.LW < 1000 Hz
Center ^{15}N linewidth	10 Hz < C.^{15}N.LW < 1000 Hz
Center ^{1}H^{N} linewidth	10 Hz < C.^{1}H^{N}.LW < 1000 Hz
Doublet1 ^{13}C linewidth	10 Hz < D1.^{13}C.LW < 1000 Hz
Doublet1 ^{15}N linewidth	10 Hz < D1.^{15}N.LW < 1000 Hz
Doublet1 ^{1}H^{N} linewidth	10 Hz < D1.^{1}H^{N}.LW < 1000 Hz
Doublet2 ^{13}C linewidth	10 Hz < D2.^{13}C.LW < 1000 Hz
Doublet2 ^{15}N linewidth	10 Hz < D2.^{15}N.LW < 1000 Hz
Doublet2 ^{1}H^{N} linewidth	10 Hz < D2.^{1}H^{N}.LW < 1000 Hz
Peak ordering	
Doublet2–Center	D2.^{13}C.PPM > C.^{13}C.PPM
Center–Doublet1	C.^{13}C.PPM > D1.^{13}C.PPM
Doublet2–Doublet1	D2.^{13}C.PPM > D1.^{13}C.PPM
Aliphatic hydrogen range	E = (D2.^{13}C.PPM – D1.^{13}C.PPM)*conv_factor– offset; lower_bound < E < upper_bound

[a] The terms for the pseudocode are defined as follows: C, center peak; D1, upfield doublet peak; D2, downfield doublet peak; P.RMSFit, Sparky's RMS Fit score for peak P; P.D.PPM, chemical shift in ppm of peak P in dimension D.; P.D.LW, linewidth in Hertz of peak P in dimension D; conv_factor, offset, lower_bound, upper_bound – RD-TR NMR experiment-specific parameters.

but not all, *Relationship Tests* are similar to statistical hypothesis tests. A typical example is a *Relationship Test* that requires the Euclidian distance between chemical shift values for two particular peaks in a group to be within a cut-off value. We define two categories of *Relationship Tests* in a *Pattern Model*; *Equivalence Relationship Tests* and *Non-Equivalence Relationship Tests* (Table 2). *Equivalence Relationship Tests* can be partitioned into separate *Equivalence Relations*. The set of *Equivalence Relationship Tests* comparing the same dimension (i.e., chemical shift value) of all peaks to each other in a *Peak Group* represents a single *Equivalence Relation*. In other words, an *Equivalence Relation* describes a set of values that are all comparable to each other. The *Equivalence Relations* provide the majority of the discriminating power of the *Pattern Model*. The number of *Equivalence Relations* is a good measure of the complexity of the *Pattern Model*.

The set of confirmed *Peak Groups* (referred to as the *Confirmed List*) provides a sampling of distances (with a mean and standard deviation) characteristic of the particular data set. These values are then used to reparameterize *Relationship Tests* in subsequent iterations. *Relationship Tests* are usually constructed so that the ideal population mean is zero and the cut-off value is based on a set number of standard deviation units. Cut-offs for *Equivalence Relationship Tests* in the same *Equivalence Relation* use the standard deviation for the *Equivalence Relation* as a whole. The initial standard deviations are guesses based upon what has worked before. *Pattern Picker* derives the standard deviation for each *Relationship Test* or *Equivalence Relation* using the *Confirmed List* and refines the value with each iteration.

Most *Relationship Tests* define a random variable $X = f(Peak\ Group)$, where f is an algebraic function that calculates the relationship of interest (e.g., the Euclidean distance between two peaks in the example above). *Pattern Picker* assumes X is approximately normally distributed. To conduct a hypothesis test, the *Peak Group* is assumed valid with the alternative hypothesis being the *Peak Group* is invalid. By default, *Pattern Pick-*

er uses a confidence level of 99.9999998% which translates into six standard deviations from the mean. If the p value of X is less than 0.0000002%, the null hypothesis is rejected and it is concluded that the *Peak Group* is invalid. Hence, the probability of Type I error, namely rejecting a real *Peak Group*, is 0.0000002%. This is a good approximation for near normal distributions. Even if the distribution is far from normal, Chebyshev's inequality[2] bounds Type I error to at most 2.78%. As the number of *Relationship Tests* increases, the number of hypothesis tests increases, and one can become more confident in the veracity of a particular *Peak Group*. *Relationship Tests* within an *Equivalence Relation* are very related and thus dependent; but *Relationship Tests* from different *Equivalence Relations* are orthogonal to each other.

Another important aspect of the *Pattern Model* is an *Allowed Missing Method* which determines whether a particular peak in a *Peak Group* can be absent. Allowing peaks in a group to be missing can account for overlapped weak or absent peaks in the spectrum and derived peak list. For example, in experiments with *Peak Groups* of size 15, having one or two missing peaks in a group will modestly reduce the number of *Relationship Tests*, and thus not reduce the confidence in a *Peak Group* significantly given that the *Relationship Tests* are equally distributed across the peaks in a group. However, in experiments with only three peaks in a group, such as many RD-TR NMR spectra, allowing a missing peak will halve the number of *Relationship Tests* and, more importantly, often eliminate a whole *Equivalence Relation* (Table 2), seriously reducing the confidence in a *Peak Group*.

Many NMR experiments exhibit overlap between peaks from different *Peak Groups*. The *Pattern Model* contains a *Consistency Method* that determines whether a *Peak Group* has an allowed type of overlap with other *Peak Groups* in the *Confirmed List*. In general, this method checks to see if a particular *Peak Group* is consistent with the *Confirmed List* (i.e., the *Peak Groups* defined to be "real"). For example, some peaks in a *Peak Group* are only allowed to belong to one group. If two *Peak Groups* both have this peak then they are not consistent.

Lastly, a *Pattern Model* includes a *Decoding Method* that extracts the information encoded in a *Peak Group*. We define (i) *additionally encoded information* as the information encoded in the target peak pattern and not directly measurable from the characteristics of a single peak, (ii) *pattern-encoding dimensions* as those dimensions of the peaks encoding the additional information, and (iii) *non-pattern-encoding dimensions* as those dimensions not encoding any additional information. For RD-TR experiments, the *additionally encoded information* is the "projected" chemical shift encoded in the in-phase splitting of the doublet peaks registered in the *pattern-encoding dimension* of the N-dimensional spectrum; e.g., the H^α chemical shift is the *additionally encoded information* in the ^{13}C *pattern-encoding dimension* in Fig. 1.

Optionally, a *Pattern Model* may have additional *Relationship Tests* used to evaluate the quality of the *Confirmed List*. A *Pattern Model* may also have an *Error Detection Method* that uses statistics calculated from the *Confirmed List* to compare related confirmed *Peak Groups* for systematic errors like sinc wiggles. A *Pattern Model* may also have an *Outlier Detection Method* for detecting abnormal outlier *Peak Groups* using the unused peaks, detected overlapped peaks, and statistics calculated from the *Confirmed List*.

2.3. Algorithm details

Fig. 3 presents a flow chart of *Pattern Picker's* main steps, while Fig. 4 shows a more detailed version of *Pattern Picker's* algorithm in the form of pseudo code. First, *Pattern Picker* selects a *Pattern Model* to use. It then reads the *Logical Peak Sets* from the input. Some NMR experiments, such as GFT NMR experiments, provide each peak in a group in a different subspectrum, while others, like some RD-TR NMR experiments (e.g., [7,30,35]), provide all peaks in the same spectrum (that is, they do not require a G-matrix transformation). Each subspectrum in an NMR experiment is represented by a *Logical Peak Set* and peak i in a *Peak Group* can only be selected from the appropriate logical set. Hence, depending on the experiment, the *Logical Peak Sets* can be pair-wise disjoint or might contain non-empty intersections.

As described above, DFS with short-circuiting creates the *Potential List* of *Peak Groups* using the *Relationship Tests* from the selected *Pattern Model*. Most *Relationship Tests* involve only a subset of the peaks in a *Peak Group*. The DFS performs the *Relationship Test* as soon as it has selected all peaks necessary for that test. The tests act as the short circuits in the DFS and inhibit exploring unfruitful paths. Fig. 2 illustrates the process by which a DFS with short circuiting finds a *Potential Peak Group* for a RD-TR NMR *Pattern Model*. Incomplete *Peak Groups* (i.e., groups with missing peaks) are created during the DFS by employing the *Allowed Missing Method* from the *Pattern Model*.

Next, each *Peak Group* in the *Potential List* is scored. The score is a likelihood that allows comparisons between different potential *Peak Groups* that conflict with each other. Usually the likelihood is a multiplication of *Equivalence Relation* probabilities, since all *Equivalence Relationship Tests* are formulated to allow easy calculation of probabilities (e.g., their

[2] $P(x - \mu \geqslant k\sigma) \leqslant 1/k^2$; where μ is the mean, σ is the standard deviation, and k is the number of standard deviation units [1].

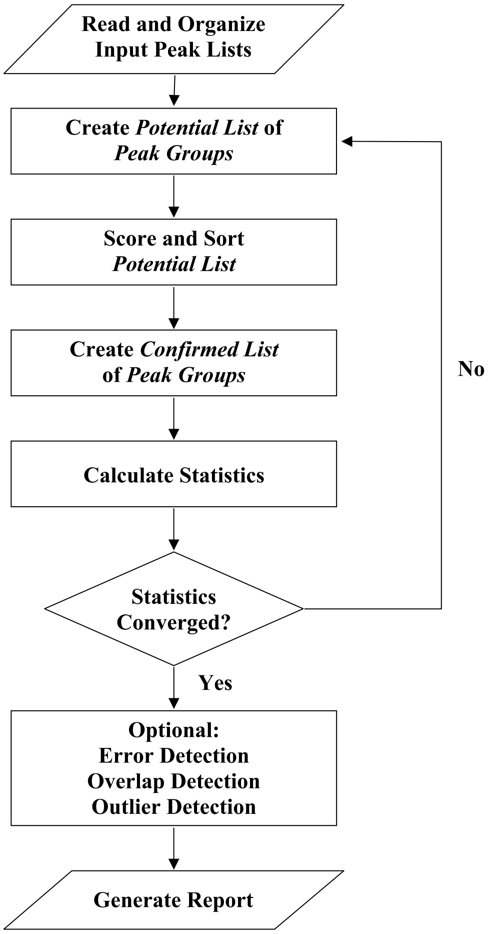

Read and Organize Input Peak Lists
Create *Potential List* of *Peak Groups*
Score and Sort *Potential List*
Create *Confirmed List* of *Peak Groups*
Calculate Statistics
Statistics Converged?
No
Yes
Optional:
Error Detection
Overlap Detection
Outlier Detection
Generate Report

Fig. 3. Flowchart for core algorithms of *Pattern Picker*.

expected mean is zero and their standard deviations calculated from previous *Confirmed Lists*). Usually, all *Non-Equivalence Relationship Tests* are excluded from the calculation of the score since they are normally constructed as simple filters and are not easily used to calculate a probability.

$$\text{Likelihood} = \prod_{i}^{N} \text{Prob}(\chi^2_{DF_i} \geqslant T_i), \qquad (1)$$

where N is the number of *Equivalence Relations*, T_i is the test statistic for *Equivalence Relation i*, and DF_i are the degrees of freedom for T_i.

A single *Equivalence Relation* probability is calculated assuming a χ^2 distribution of DF_i degrees of freedom. The test statistic is the summation of DF_i "worst" *Equivalence Relationship Tests* divided by their corresponding standard deviation (Eq. (2)). Each "worst" *Equivalence Relationship Test* is an independent variable corresponding to the worst value from a row of related *Equivalence Relationship Tests* in a matrix representing an *Equivalence Relation*.

$$T_i = \sum_{j}^{DF_i} \frac{(X_j)^2}{\sigma_i^2}, \qquad (2)$$

where DF_i is the degrees of freedom for T_i, X_j are the worst random variable result from set j (matrix row) of related *Equivalence Relationship Tests*, and σ_i is the standard deviation for T_i (*Equivalence Relation i*).

Since the *Equivalence Relation* probabilities are not framed as probabilities of the *Peak Group*'s existence, their multiplication in Eq. (1) yields a likelihood and not a probability; however, in practice this score works well since it is only used to rank *Peak Groups* relative to each other.

The *Potential List* is then sorted, in descending order, by the likelihood score of each *Peak Group* (Eq. (1)). Next, *Pattern Picker* moves *Peak Groups* from the *Potential List* to the *Confirmed List*, one by one in a best first approach according to their scores. But before moving a *Peak Group* to the *Confirmed List*, *Pattern Picker* uses the *Consistency Method* from the selected *Pattern Model* to ensure that the new *Peak Group* is consistent with preexisting *Peak Groups* in the *Confirmed List*. Since *Peak Groups* with higher scores are added first and incorrect groups should theoretically have lower likelihood scores, only correct *Peak Groups* should be added to the *Confirmed List*. This best first approach to verifying *Peak Groups* is faster than a global optimization approach and has the added benefit of isolating confidence or reliability of individual parts of the results.

The only constraint on the DFS is that *Peak Groups* have to fit the *Pattern Model* defined for the particular NMR experiment. Hence, the DFS can create two *Peak Groups* that fit the *Pattern Model*, but are inconsistent with each other. For example, peaks from a restricted *Logical Peak Set* are allowed to be members of only one *Peak Group*; however, the DFS can create two groups that share a particular peak from that restricted *Logical Peak Set*. The probability of putting an invalid *Peak Group* (i.e., Type II error related to the false positive rate) into the *Potential List* is significantly higher than the probability of rejecting a valid *Peak Group* (i.e., Type I error ~ 0.000001% related to the false negative rate) because Types I and II errors are inversely related when the sample size is constant [29]. However, invalid *Peak Groups* in the *Potential List* will conflict with valid *Peak Groups* in the *Confirmed List*. Thus, *Pattern Picker* is simultaneously sensitive and selective enough to keep both the false negative and false positive rates low.

The *Confirmed List* of *Peak Groups* contains the groups that *Pattern Picker*, believes are real *Peak Groups*. However, the initial standard deviations used in the *Relationship Tests* to determine these *Peak Groups* are based on typical values observed in previously analyzed spectra and do not reflect unique features of the particular spectrum under analysis. As such, this preliminary *Confirmed List* is used to calculate a new set of standard deviations for the *Relationship Tests*. The standard deviations are stored, and the entire preceding

```
1. Select Pattern Model to use (based on experiment)
2. Parse Peak lists into logical sets of peaks.
3. For Peak Set A = 1 to N
   A) For each Peak X in set A (or no peak selected if method in Pattern Model indicates
      that a valid pattern may lack this peak)
      I. If Peak X passes all Relationship Tests within given tolerances, then select it and
         goto Step C
   B) Next X
   C) Decrement A by 2 and goto Step A (i.e. Select another peak from the previous set)
   D) If A=N and the group of selected peaks passes all Relationship Tests, create a
      potential Peak Group and add it to the Potential List of Peak Groups. Decrement
      A by 2 (i.e. Select another peak from the previous peak set)
4. Next A
5. Score each potential Peak Group based on results of the Equivalence Relationship
   Tests
6. Sort the Potential List in descending order of the score
7. For each potential Peak Group Z in the sorted list
   A) Use the Consistency Method in the Pattern Model to ensure that Z is consistent
      with the Confirmed List of Peak Groups
   B) If Z is consistent, add it to the Confirmed List and remove any sub-groups of Z
      (incomplete Peak Groups that are equivalent to Z).
8. Next Z
9. Calculate new standard deviations for Relationship Tests using the Confirmed List.
10. Compare these standard deviations to previously derived standard deviations
11. If standard deviation comparison fails the convergence test and the number of
    iterations has not reached the maximum goto Step 3.
12. Perform Error Detection (Optional).
13. Perform Peak Overlap Detection (Optional).
14. Perform Outlier Peak Group Detection (Optional).
15. Generate Report: Run each Peak Group through additional Relationship Tests as well
    as calculate additional statistical information.
16. Output Results
```

Fig. 4. Pseudo code for core algorithms of *Pattern Picker*.

algorithm is re-run with these new statistics. The benefit of this approach is that the calculated standard deviations are better estimators for their respective populations than ones provided a priori. Also, better standard deviations create better cut-offs for the *Relationship Tests* and more accurate χ^2 probabilities for the scoring function. Both adjustments improve *Pattern Picker's* ability to identify legitimate *Peak Groups* and distinguish "real" from "false" peaks.

When the preceding algorithm is run the next time, the *Confirmed List* will generally change and more accurately reflect the true *Peak Groups* in the data set. Since the two *Confirmed Lists* are different, the calculated standard deviations for the previous *Confirmed List* will be different. The two sets of calculated standard deviations are compared using an approximation to the *F* test to test whether the sample statistics are a good measure of the population statistics. If any standard deviation calculated from the subsequent run is significantly different from a deviation calculated in a prior run, then the affected *Relationship Tests* will not accurately discriminate valid *Peak Groups* from invalid *Peak Groups*. This could affect either the false negative or false positive rates of the *Confirmed List* in this run. In this case, the algorithm repeats with the new set of standard deviations. This process iterates until the set of calculated standard deviations from the ith *Confirmed List* converge, i.e., they are not significantly different from the standard deviations calculated for the $i - 1$st *Confirmed List*. As long as the data is near normal and the early *Confirmed Lists* have a significant proportion of real *Peak Groups*, this iterative process generally converges well either upwards from too few *Peak Groups* in the *Confirmed List* or downwards from too many *Peak Groups* in the *Confirmed List*. In practice, upward convergence is faster since fewer *Peak Groups* are involved and those present in the early *Confirmed Lists* are generally completely correct. In tests carried out to date, almost every data set has eventually converged within 15 iterations. The iterative process ensures that the standard deviations used by the *Relationship Tests* accurately reflect the data. In addition, they remove the uncertainty for users by not requiring them to enter any critical numerical parameters.

After convergence, a DFS may be run on the remaining unused peaks, and detected overlapped peaks with looser constraints to find any outlying *Peak Groups* (*Outlier Detection Method*). The results of the outlier detection run are less accurate and generally require manual culling. However, they serve as a useful guide for identifying less well-defined (outlier) *Peak Groups*.

Finally, *Pattern Picker* generates a summary report detailing information and statistics on each *Peak Group*, including the constituent peaks and the likelihood score for the group. The results of any additional *Relationship Tests* not used in creating the *Potential Lists* are also included to aid in error analysis. These additional *Relationship Tests* are not as reliable as the ones used to create the *Confirmed List*; however, they provide a useful validation of the selected *Peak Groups*. Also, the results from the *Error Detection Method* and *Outlier Detection Method* in the *Pattern Model* are included as well as statistics on detected and/or suspected overlapped peaks. The summary report facilitates the verification of the results and serves as an aid for additional manual analysis of spectra if such analysis is needed.

2.4. Reduced-dimensionality triple resonance NMR pattern model

In this paper, we apply the general peak editing algorithm described above in the analysis of RD-TR NMR spectra; applications to other kinds of NMR data will be presented elsewhere. RD-TR NMR experiments provide spectral data as a group of three peaks. As described above, an *N*-dimensional RD-TR NMR experiment encodes $N + 1$ chemical dimensions in each *Peak Group*. The *Equivalence Relations* for this pattern are displayed pictorially in Fig. 1. Each peak in a *Peak Group* shares the same chemical shift for $N - 1$ *non-pattern-encoding dimensions*. For the *N*th *pattern-encoding dimension*, each peak in the *Peak Group* has a distinct chemical shift: the central peak exhibits the chemical shift in the *N*th *pattern-encoding dimension* about which the outer doublet peaks are centered. The difference of the shifts of the outer peaks is proportional to the chemical shift of an $N + 1$st "projected" dimension and represents the *additionally encoded information*.

The complete set of *Relationship Tests* for the RD-TR NMR *Pattern Model* is summarized in Table 2. The ^{15}N-Matching and ^{1}H^{N}-Matching *Equivalence Relations* ensure that every peak in a *Peak Group* has similar chemical shift values in the *non-pattern-encoding dimensions*. A Doublet Symmetry *Equivalence Relation* ensures that the central peak is the midpoint of the doublet peaks in the *pattern-encoding dimension*. All of the *Relationship Tests* comprising these three *Equivalence Relations* are hypothesis tests. Not all *Relationship Tests* in the RD-TR NMR *Pattern Model* (Table 2) are hypothesis tests and members of an *Equivalence Relation*. These include general peak filtering *Relationship Tests* that filters obvious "false" peaks. In this work, we used the program Sparky [10] for automated peak picking. Sparky provides peak quality assessment scores (i.e., Fit RMS and linewidth), which are helpful in identifying candidate "false" peaks. Sparky's Fit RMS score is a weighted root mean square of the deviations of a peak from an ideal Lorentzian lineshape [10]. The RD-TR *Pattern Model* includes a *Non-Equivalence Relationship Test* that only accepts peaks with Fit RMS score between 0 and 500. Another *Non-Equivalence Relationship Test* only accepts peaks with linewidths between 10 and 1000 Hz. The upper bound is set unrealistically high due to occasional inaccuracies in line widths reported by Sparky. The RD-TR *Pattern Model* also includes a range test which requires the encoded chemical shift value to be within an acceptable range. For instance, if the encoded shift is an aliphatic hydrogen, and a particular *Peak Group* projected a value outside the typical range of -2 to 7 ppm, then that particular *Peak Group* would be suspect. The last *Non-Equivalence Relationship Test* is a peak-ordering test that requires the central peak to be between the doublet peaks in the *pattern-encoding dimension*.

The likelihood scoring function for the RD-TR *Pattern Model* is based on the *p* values for the matching of each peak in a *Peak Group* in the *non-pattern-encoding dimensions* and the *p* value for the symmetry of the central peak between the two doublet peaks. Since RD NMR *Peak Groups* are only composed of three peaks, permitting one missing peak drastically reduces the number of testable *Relationship Tests*, especially the Doublet Symmetry *Relationship Test*. Thus, while potentially accommodated in *Pattern Models* of other types of NMR experiments, missing peaks are not allowed in the RD-TR *Pattern Model*. The *Consistency Method* ensures that a doublet peak is not a member of two different *Peak Groups*. In some RD-TR NMR experiments, reusing the central peak is permitted, allowing for resonance degeneracy in the *non-pattern-encoding* and *pattern-encoding dimensions*.

A group of noise (random false) peaks successfully passing all the RD-TR *Relationship Tests* is very improbable. However, systematic anomalies can be a significant issue, such as sinc wiggles that create false peaks around a legitimate peak. If each peak in a group has sinc wiggles as shown in Fig. 5, then those sets of false peaks can form an invalid *Peak Group* that passes all the RD *Relationship Tests*.

Invalid *Peak Groups* due to sinc wiggles will always be near a legitimate *Peak Group*. In addition, the peaks in a group due to sinc wiggles will have a much weaker intensity than the peaks in the legitimate group. The Doublet Symmetry *Relationship Test* ensures that the invalid *Peak Group* will also have a chemical shift values similar to those of a legitimate *Peak Group*. Thus, to detect invalid groups due to sinc wiggles, *Pattern Picker* searches for *Peak Groups* with similar chemical shifts in the *non-pattern-encoding dimensions* and a similar chemical shift in either the *pattern-encoding dimension*

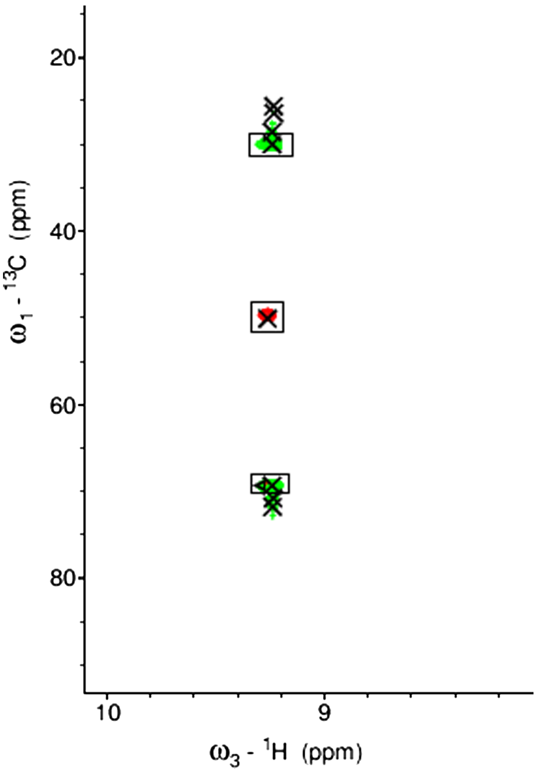


Fig. 5. Sinc wiggles in the HABCABcoNH spectrum of *E. coli* yggU. The real peaks have boxes drawn around them, while sinc wiggles do not. The sinc wiggles around the doublet peaks are symmetric around the central peak and thus pass all the *Relationship Tests* in the RD-TR NMR *Pattern Model* (Fig. 1). However, the *Error Detection Method*, discussed in the text, detects them. While it is possible to suppress sinc wiggles for a specific peak by appropriate adjustment of processing window functions, in general the "best" window function for a particular spectrum will be suboptimal for the sharpest peaks in the spectrum, and will result in sinc wiggles for some peaks.

(i.e., the chemical shift of the central peak) or the *additionally encoded information* (i.e., the encoded chemical shift value). The average intensities of two *Peak Groups* that fit these criteria are compared and if one *Peak Group* has an average intensity that is half an order of magnitude less than the other group then the weaker group is marked as a potential sinc wiggle artifact. These "sinc wiggle" *Peak Groups* are left in the output, but are marked as likely to be spectral artifacts.

Another method for error detection is a simple count of the number of weak peaks in a *Peak Group*. The input for this analysis includes Sparky's Fit RMS score. If this Fit RMS score is low, a peak is considered weak. Although we expect a certain number of weak peaks in any experiment, a *Peak Group* with multiple weak peaks is considered suspect.

The *Consistency Method* for the RD-TR NMR *Pattern Model* permits reuse of central peaks; however, it does not allow doublet peaks to belong to two different *Peak Groups*. In certain cases, doublet peaks in an RD-TR NMR experiment overlap. For instance, in a HABCABcoNH experiment, the α and the β carbon cross peaks fall in a line defined by their common ^{15}N and ^{1}H^N frequencies. Occasionally, the upfield component of a doublet around an α carbon will overlap with the downfield component of a doublet around the β-carbon. In this case, there will only be three doublet peaks in the peak list when four are expected. Although both legitimate *Peak Groups* will be created by the DFS, one group will initially be eliminated as being inconsistent with the other *Peak Group* during the consistency check. This is the appropriate behavior because in the majority of cases this eliminates invalid *Peak Groups* formed with some legitimate peaks and some false peaks. However in this case, this is not the correct behavior because both *Peak Groups* are legitimate.

To detect these cases of overlap, we first assume that the H$^\alpha$–C$^\alpha$ *Peak Group* was added to the *Confirmed List* and the H$^\beta$–C$^\beta$ *Peak Group* was rejected by the consistency check. The upfield peak of the H$^\alpha$–C$^\alpha$ *Peak Group*, which overlaps with the downfield peak of the H$^\beta$–C$^\beta$ *Peak Group*, will have a much stronger intensity than the non-overlapped downfield peak of the H$^\alpha$–C$^\alpha$ *Peak Group*. With the assumption that the differences in intensity are normally distributed, with a mean of zero and a standard deviation derived from the data, *Pattern Picker* can thus detect that the H$^\alpha$–C$^\alpha$ *Peak Group* includes one overlapped peak. These overlapped peaks are indicated in the report. They are then added to the list of unused peaks and used in an outlier *Peak Group* detection run to automatically identify the H$^\beta$–C$^\beta$ *Peak Group*.

3. Results

Escherichia coli open reading frame yggU codes for a 108 residue protein with unknown function (Swiss-Pro ID, P52060). NMR data for yggU were collected at 20 °C on a four-channel Varian INOVA 600 MHz NMR spectrometer. The data were processed with NMRPipe v2.1 [6]. Table 3 shows the acquisition and processing parameters for HACAcoNH and HABCABcoNH RD-TR NMR experiments used in testing *Pattern Picker*.

After processing, we aligned each of the HACAcoNH and HABCABcoNH spectra with a manually peak-picked ^{15}N-edited HSQC spectrum. We then used the spectral visualization software Sparky [10] to restrictively peak pick these spectra using tolerances of ±0.04 and ±0.4 ppm in the ^{1}H^N and ^{15}N dimensions, respectively. The restrictive peak lists were then filtered by intensity to include twice as many peaks as expected in the experiment. These intensity-filtered restrictive peak lists are denoted as "raw" peak lists throughout the paper. Manual inspection of both the HACAcoNH and HABCABcoNH spectra showed additional correlations arising from two unassigned histidines in the N-terminal hexa-histidine (hexa-His) tag used for affinity purifica-

Table 3
Acquisition and processing parameters for RD-TR HACAcoNH and HABCABcoNH spectra of *E. coli* yggU

Parameters	HACAcoNH	HABCABcoNH
Acquisition		
Frequency labeling (dimension)	$C(\omega 1)$ $N(\omega 2)$ $H(\omega 3)$	$C(\omega 1)$ $N(\omega 2)$ $H(\omega 3)$
Data collection size (points)	95 × 32 × 512	95 × 32 × 512
Number of transients	1	2
sw (Hz)	14998, 2123, 8384	14998, 2123, 8384
$T_{1,\ \max}$ (ms)	6.3, 15.1, 61.1	6.3, 15.1, 61.1
Recycle delay (s)	1.06	1.06
Total time (h)	3.9 × 2	8 × 2
Processing		
Final size (points)	512 × 256 × 1024	512 × 256 × 1024
Window function	Sine bell	Sine bell
Linear prediction	In $\omega 2$	In $\omega 2$

tion of yggU. These correlations from the flexible tag were not observed in the ^{15}N-edited HSQC and, apparently, are enhanced by a relaxation-filtering effect in the RD-TR NMR experiments. Thus, the automatically peak-picked raw peak lists, which have been filtered to remove peaks not represented in the ^{15}N-edited HSQC spectrum, lack peaks for these two histidines. We also filtered these raw peak lists against a set of manually peak-picked lists available for these same data sets [3] and removed matching peaks. We then concatenated the resulting "false" peaks with the manual lists to create the "manual + false" peak lists. Analyses were thus carried out for manually curated ("manual"), "manual + false," and "raw" peak lists. The "raw" peak lists represent the typical RD-TR NMR data available without manual analysis and curation, whereas the "manual + false" peak lists represent complete RD-TR NMR data with real "false" peaks.

Results of *Pattern Picker* using peak lists from RD-TR HACAcoNH and HABCABcoNH spectra of yggU are presented in Tables 4 and 5, respectively. In these analyses, *accuracy* is defined as the number of correct *Peak Groups Pattern Picker* recognized divided by the total number of *Peak Groups* identified by the program. "False positives" are incorrect *Peak Groups* that *Pattern Picker* identified, and "yield" is the number of correct *Peak Groups* recognized by the program divided by the number of *Peak Groups* determined by a manual analysis. "False negatives" are valid *Peak Groups* that *Pattern Picker* did not recognize.

Pattern Picker was tested with the manual peak lists, raw peak lists, and manual + false peak lists, for each experiment described above. Tests with the manual peak lists demonstrate the ability of *Pattern Picker's* core algorithms and methodology to properly group peaks together. For instance, in the HABCABcoNH, it is important that *Pattern Picker* does not confuse a doublet belonging to a β carbon for a doublet belonging to an α carbon. Tests with the manual + false peak list measure *Pattern Picker's* ability to distinguish legitimate groups from invalid groups when realistic false peaks are included. These tests are independent of the data completeness since all real peaks are present in the peak lists. Tests with the raw peak lists measure *Pattern Picker's* ability to fully automate the peak picking/editing analysis of RD-TR NMR experiments.

For both the HACAcoNH and HABCABcoNH datasets, *Pattern Picker* obtains 100% accuracy and 0% false positives on all three peak lists; i.e., all of the *Peak Groups* identified by the program are true *Peak Groups*. The

Table 4
Results of RD-TR HACAcoNH data analysis for *E. coli* yggU protein

Input list	Accuracy	False positives	Yield	False negatives
Manual	100% (102/102)	0% (0/102)	99.0% (102/103)	0.97% (1/103)
Raw	100% (99/99)	0% (0/99)	96.1%[a] (99/103)	3.88%[b] (4/103)
Manual + False	100% (103/103)	0% (0/103)	100% (103/103)	0% (0/103)

[a] This yield does not reflect the fact that the manually peak picked HSQC used for restrictive peak picking did not have peaks corresponding to two correlations in the HACAcoNH spectrum of *E. coli* yggU arising from the hexa-His tag, as explained in the text. The adjusted yield is 98.0% (99/101).

[b] This false negative value does not reflect the fact that the manually peak picked HSQC used for restrictive peak picking did not have peaks corresponding to two correlations arising from the hexa-His tag in the HACAcoNH spectrum of *E. coli* yggU. The adjusted false negative value is 1.98% (2/101).

Table 5
Results of RD-TR HABCABcoNH data analysis for *E. coli* yggU protein

Input list	Accuracy	False positives	Yield	False negatives
Manual	100% (203/203)	0% (0/203)	99.0% (203/205)	0.98% (2/205)
Raw	100% (182/182)	0% (0/182)	88.8%[a] (182/205)	11.2%[b] (23/205)
Manual + False	100% (205/205)	0% (0/205)	100% (205/205)	0% (0/205)

[a] This yield does not reflect the fact that the manually peak picked HSQC used for restrictive peak picking did not have peaks corresponding to four correlations in the HABCABcoNH spectrum of *E. coli* yggU arising from the hexa-His tag as explained in the text. The adjusted yield is 90.5% (182/201).

[b] This false negative value does not reflect the fact that the manually peak picked HSQC used for restrictive peak picking did not have peaks corresponding to four correlations arising from the hexa-His tag in the HABCABcoNH spectrum of *E. coli* yggU. The adjusted false negative value is 9.45% (19/201).

program obtains nearly 100% yield on manual and manual + false peak lists, and 96.1 and 88.8% yield on the raw peak lists in the HACAcoNH and HABCABcoNH experiments respectively; however, these numbers do not reflect the fact that the manually peak picked ^{15}N-edited HSQC peak list used in the restrictive peak picking was missing two peaks corresponding to two *Peak Groups* in the HACAcoNH spectrum and four *Peak Groups* in the HABCABcoNH spectrum. These *Peak Groups* correspond to two unassigned histidine residues in the hexa-His tag of the protein sample as explained above. Considering that these peaks are eliminated from analysis by the restrictive peak picking, the adjusted yields are 98.0 and 90.5% in the HACAcoNH and HABCABcoNH experiments, respectively. With these data sets, the fully automated peak editing of *Pattern Picker* performs almost as well as human manual analysis. During error detection of the HACAcoNH *Confirmed List*, *Pattern Picker* detected all five

Table 6
Results for different restrictive peak picking protocols

Input list	% of Expected	Accuracy	False positive[a]	Yield	False negative
HACAcoNH Raw Peak List					
0.04/0.4 RPP[f], Intensity filtered	200	100% (99/99)	0% (0/99)	96.1%[b] (99/103)	3.88%[d] (4/103)
0.04/0.4 RPP[f]	285	100% (101/101)	0% (0/103)	98.1%[b] (101/103)	1.94%[d] (2/103)
0.06/0.6 RPP[g]	404	100% (101/101)	0% (0/101)	98.1%[b] (101/103)	1.94%[d] (2/103)
HABCABcoNH Raw Peak List					
0.04/0.4 RPP[f], Intensity filtered	200	100% (182/182)	0% (0/182)	88.8%[c] (182/205)	11.2%[e] (23/205)
0.04/0.4 RPP[f]	265	100% (185/185)	0% (0/185)	90.2%[c] (185/205)	9.76%[e] (20/205)
0.06/0.6 RPP[g]	383	98.4% (190/193)	1.55% (3/193)	92.7%[c] (190/205)	7.32%[e] (15/205)

[a] All errors have the worst scores and are at the bottom of the *Confirmed List* once sinc wiggles errors have been removed. They fail most of the additional tests shown in the report.

[b] These yields do not reflect the fact that the manually peak picked HSQC used for restrictive peak picking did not have peaks corresponding to two correlations arising from the hexa-His tag in the HACAcoNH spectrum of *E. coli* yggU. The adjusted yields are 97.0% (99/101), 100% (101/101), and 100% (101/101).

[c] These yields does not reflect the fact that the manually peak picked HSQC used for restrictive peak picking did not have peaks corresponding to four correlations arising from the hexa-His tag in the HABCABcoNH spectrum of *E. coli* yggU. The adjusted yields are 90.5% (182/201), 92.0% (185/201), and 94.5% (190/201).

[d] These false negative values do not reflect the fact that the manually peak picked HSQC used for restrictive peak picking did not have peaks corresponding to two correlations arising from the hexa-His tag in the HACAcoNH spectrum of *E. coli* yggU. The adjusted false negative values are 1.98% (2/101), 0% (0/101), and 0% (0/101).

[e] These false negative values do not reflect the fact that the manually peak picked HSQC used for restrictive peak picking did not have peaks corresponding to four correlations arising from the hexa-His tag in the HABCABcoNH spectrum of *E. coli* yggU. The adjusted false negative values are 9.45% (19/201), 7.96% (16/201), and 5.47% (11/201).

[f] Restrictive peak picking with tolerances of ±0.04 and ±0.4 ppm in the ^{1}H^{N} and ^{15}N dimensions, respectively.

[g] Restrictive peak picking with tolerances of±0.06 and±0.6ppm in the ^{1}H^{N} and ^{15}N dimensions, respectively.

Peak Groups due to sinc wiggles from the manual + false peak list and all four *Peak Groups* due to sinc wiggles present in the raw peak list. During error detection of the HABCABcoNH *Confirmed List*, the program detected all six *Peak Groups* due to sinc wiggles present in both the manual + false peak list and the raw peak lists. The sinc wiggle error detection exhibited no false positives in analyzing these spectra.

In Table 6, we show results obtained when we vary the restrictive peak picking protocol of Sparky. When we increase the number of artifactual peaks in the input peak lists by expanding the restrictive peak picking tolerances to ±0.06 and ±0.6 ppm in the $^{1}H^{N}$ and ^{15}N dimensions, respectively, the yield for the HACAcoNH raw lists improves slightly from 96.1% (99/103) to 98.1% (101/103) with no change in accuracy; when the missing hexa-His tag peaks from the manual HSQC are considered, the calculated yield is 100% (101/101). For the HABCABcoNH raw lists, the less conservative peak picking increases the yield from 88.8% (182/205) to 92.7% (190/205); i.e., 94.5% (190/201), once the missing peaks from the manual HSQC list are considered. However, the accuracy drops to 98.4% (190/193) with a 1.6% (3/193) false positive rate; but, the three incorrectly identified *Peak Groups* are among the seven with lowest scores in the *Confirmed List*, and fail some of the additional tests in the report. Overall, these results demonstrate the robustness of the *Pattern Picker* algorithm in distinguishing real from artifactual peaks in the input peak lists.

4. Discussion

The excellent and robust results of *Pattern Picker* on manual and manual + false peak lists demonstrate the discriminating power of the RD-TR NMR *Pattern Model* to detect correct *Peak Groups* in these types of RD-TR NMR experiments. Most of the discrimination power arises from the three *Equivalence Relations* of *Relationship Tests*: ^{15}N-Matching, $^{1}H^{N}$-Matching, and Doublet Symmetry (Table 2). Table 7 shows the discriminating power of each set of *Relationship Tests* separately and all combinations of *Equivalence Relations* on the raw peak lists. The numbers represent the size of the *Potential List* of *Peak Groups* when using only the given *Relationship Tests*. The numbers in parenthesis are the percent ratio to a *Potential List* when no *Relationship Tests* are used. When no *Relationship Tests* are used to discriminate, the sizes of the *Potential Lists* are 14,384,898 and 165,487,086 *Peak Groups*, for the HACAcoNH and HABCABcoNH raw peak lists, respectively. Of the three *Equivalence Relations*, the ^{15}N-Matching *Equivalence Relation* is the most discriminating. This is to be expected since the dispersal of chemical shift values is larger for the ^{15}N-Matching

Table 7
Relationship test discrimination power

Relationship Tests Used	yggU HACAcoNH[a]	yggU HABCABcoNH[b]
Non-Equivalence Relationship Tests		
Peak Filtering	4,949,490 (34.4%)	39,991,830 (24.2%)
Peak Ordering	2,640,372 (18.4%)	28,815,157 (17.4%)
Aliphatic hydrogen range	6,786,738 (47.2%)	117,302,525 (70.9%)
Equivalence Relations		
^{15}N-Matching	4658 (0.0324%)	119,273 (0.0721%)
$^{1}H^{N}$-Matching	24,603 (0.171%)	538,592 (0.3255%)
Doublet symmetry	143,538 (0.100%)	1,616,654 (0.9769%)
^{15}N-Matching and $^{1}H^{N}$-Matching	1284 (0.00890%)	21,149 (0.0128%)
^{15}N-Matching and Doublet symmetry	162 (0.00110%)	1467 (0.000886%)
$^{1}H^{N}$-Matching and Doublet symmetry	382 (0.00270%)	5604 (0.00339%)
All Equivalence Relations	122 (0.000800%)	452 (0.000273%)
All Relationship Tests	111 (0.000772%)	241 (0.000146%)

[a] The total number of combinations of three peaks from the raw peak list: 14,384,898.

[b] The total number of combinations of three peaks from the raw peak list: 165,487,086.

Equivalence Relation than for the $^{1}H^{N}$-Matching and Doublet Symmetry *Equivalence Relations*. However, the discriminating power of combinations of *Equivalence Relations* is not as intuitive. The ^{15}N-Matching and Doublet Symmetry combination is almost eight times more discriminating than the ^{15}N-Matching and $^{1}H^{N}$-Matching combination in the HACAcoNH results and more than 14 times as discriminating in the HABCABcoNH results, even though the latter combination use the two *Equivalence Relations* with the best single discrimination power. The combination of all three *Equivalence Relations* along with other "common sense" *Non-Equivalence Relationship Tests* produces the *Potential List* of *Peak Groups* that *Pattern Picker* finds. The *Non-Equivalence Relationship Tests* are not very discriminating by themselves; however, for the HABCABcoNH, they almost double the overall discrimination power when combined with the three *Equivalence Relations*. The difference between the three individual *Equivalence Relation* results and all *Relationship Test* results are not that significant for the HACAcoNH raw list. The final discrimination of the *Potential List* to the *Confirmed List* arises from the best first consistency testing.

The *Pattern Picker* algorithm is very robust with respect to noise and other artifacts. It even handles very "noisy" peak lists with up to four times the number of expected peaks, with the accuracy dropping only to 98.4% (Table 6). However, it is easy to see from Table 7 why missing a single peak in a *Peak Group* greatly reduces the discriminating power of the *Pattern Model*. Missing any single peak takes the Doublet Symmetry *Equivalence Relation* away. This reduces the discriminating power by ~10-fold in the HACAcoNH raw peak list and by ~50-fold in the HABCABcoNH raw peak list. Since the RD-TR NMR *Pattern Model* is very sensitive to missing peaks, this particular *Pattern Model* requires all *Peak Groups* to be complete.

The excellent results of *Pattern Picker* with the raw peak lists demonstrate that this approach to automated peak list editing is practical. It is also convenient to use since all critical numerical parameters are derived from the data and iteratively optimized by the program. Because the RD-TR NMR *Pattern Model* is so discriminating and can handle large numbers of false peaks, weak *Peak Groups* can be detected because the underlying restrictive peak picking can cut very close to the noise level to minimize false negatives. Future development of peak pickers for use with peak list editors like *Pattern Picker* could focus on minimizing the rate of false negatives without worrying how much this increases the rate of false positives.

Still, the raw peak lists used in these analyses are not completely unedited. They are the results from restrictive peak picking using a manually peak picked ^{15}N-edited HSQC. Requiring a manually edited and inspected HSQC peak list for the process of automatic peak picking/editing of a series of RD-TR NMR spectra is reasonable since it is rather quick and easy to obtain and is generally used anyway to restrictively peak pick prior to automated NMR data analysis [18,36].

Some RD NMR experiments cannot be as effectively restrictive peak picked. For example, the RD HCCH COSY and TOCSY experiments [35] are not as cleanly restrictive peak picked using a ^{13}C-edited HSQC because of the ^{1}H–^{13}C cross peak overlap in this 2D spectrum [18]. However, these experiments have symmetry correlations that may be used in a similar manner to the restrictive peak picking. A solvent line exclusion *Relationship Test* may be used in the core *Pattern Model* and then replaced with a symmetry correlation *Relationship Test* in the outlier detection run. Such a symmetry correlation *Relationship Test* would require the presence of a matching symmetry correlation in the *Confirmed List*.

There are other types of experiments that encode information in groups of peaks. Certain residual dipolar coupling experiments [25] and ^{1}J-resolved E-COSY experiments [19,20] have characteristic patterns, which can be supplemented and enriched by appropriate data combinations. For example, by combining aligned and non-aligned spectra, together with fully decoupled spectra, standard RDC experiments can be represented as 5-peak patterns which are highly amenable to *Pattern Picker* analysis (unpublished results). Combinations of experimental approaches can be used to create new experiments with even larger patterns. For example, the recently described G-matrix FT experiment provides a rich 15-peak pattern [15]. *Pattern Models* to handle these experiments have more *Equivalence Relations* to test and significantly more discriminating power than the RD-TR NMR *Pattern Model* presented here. With increasing discriminating power, *Pattern Picker's* results for those experiments would be even more reliable and should detect even weaker correlations in their spectra.

Pattern Picker's algorithm is also fast. A brute force approach would perform approximately 244 million *Relationship Tests* in the analysis of the yggU RD-TR HACAcoNH raw peak list. *Pattern Picker's* DFS with short-circuiting algorithm performs the same analysis with less than 2.5 million *Relationship Tests*. Each *Pattern Picker* analysis performed for this paper took less than 2s on a 1.4GHz AMD Athlon processor running RedHat Linux 7.3, except the HABCABcoNH 0.06/0.6 RPP raw analysis (Table 6), which took 8s. As the number of testable *Peak Groups* and *Relationship Tests* increases exponentially with the number of peaks in a pattern, the enhancement in performance using DFS with short-circuiting over brute force tree searches will be even larger in analyzing spectra with more complex patterns.

In conclusion, *Pattern Picker's* overall algorithm is robust on these RD-TR NMR experiments. It is also fast and practical to use. The algorithm neatly divides correlation detection into two separate problems, peak picking and peak editing, which can be separately optimized. Moreover, the program naturally fits into any automated process for peak picking and peak list editing. The program also provides facilities for handling experiments that encode information in groups of peaks, thus promoting the use of these experiments that are typically harder for a human to analyze. This opens up additional areas to explore in experiment design since the complexity of the pattern is now a benefit, and not a drawback, to analysis.

Acknowledgments

We thank Jeff Mills for acquiring the RD-TR NMR experimental data and David Snyder for helpful discussions and comments on the manuscript. This work was supported by grants from the National Science Foundation (MCB-9407569 and DBI-9974200 Postdoctoral fellowship to H.N.B.M.) and the National Institutes of Health (P50 GM62413). Nadeem Riaz was a Henry Rutgers Undergraduate Scholar. The *Pattern Picker* program is available by request from the authors.

References

[1] M. Abramowitz, I.A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, Dover, New York, 1972.

[2] M. Andrec, J.H. Prestegard, J. Magn. Reson. 130 (1998) 217–232.

[3] J.M. Aramini, J.L. Mills, R. Xiao, T.B. Acton, M.J. Wu, T. Szyperski, G.T. Montelione, J. Biomol. NMR 27 (2003) 285–286.

[4] S.E. Brenner, Nat. Rev. Gen. 2 (2001) 801–809.

[5] B. Brutscher, J. Simorre, M.S. Caffrey, D. Marion, J. Magn. Res. B 105 (1994) 77–82.

[6] F. Delaglio, S. Grzesiek, G.W. Vuister, G. Zhu, J. Pfeifer, A. Bax, J. Biomol. NMR 6 (1995) 277–293.

[7] K. Ding, A.M. Gronenborn, J. Magn. Reson. 156 (2002) 262–268.

[8] C. Eccles, P. Güntert, M. Billeter, K. Wüthrich, J. Biomol. NMR 1 (1991) 111–130.

[9] D.S. Garrett, R. Powers, A.M. Gronenborn, G.M. Clore, J. Magn. Res. 95 (1991) 214–220.

[10] T.D. Goddard, D.G. Kneller, SPARKY 3, University of California, San Francisco, 1999.

[11] T. Herrmann, P. Güntert, K. Wüthrich, J. Mol. Biol. 319 (2002) 209–227.

[12] T. Herrmann, P. Güntert, K. Wüthrich, J. Biomol. NMR 24 (2002) 171–189.

[13] Y.P. Huang, G.V.T. Swapna, P.K. Rajan, H. Ke, B. Xia, K. Shukla, M. Inouye, G.T. Montelione, J. Mol. Biol. 327 (2003) 521–536.

[14] M.A. Kennedy, G.T. Montelione, C.H. Arrowsmith, J.L. Markley, J. Struct. Funct. Genomics 2 (2002) 155–169.

[15] S. Kim, T. Szyperski, J. Am. Chem. Soc. 125 (2003) 1385–1393.

[16] S. Kim, T. Szyperski, J. Biomol. NMR 28 (2004) 117–130.

[17] R. Koradi, M. Billeter, M. Engeli, P. Güntert, K. Wüthrich, J. Magn. Reson. 135 (1998) 288–297.

[18] D. Monleon, K. Colson, H.N.B. Moseley, C. Anklin, R. Oswald, T. Szyperski, G.T. Montelione, J. Struct. Funct. Genomics 2 (2002) 93–101.

[19] G.T. Montelione, S.D. Emerson, B.A. Lyons, Biopolymers 32 (1992) 327–334.

[20] G.T. Montelione, M. Winkler, E. Rinderknecht, G. Wagner, J. Magn. Reson. 82 (1989) 198–204.

[21] G.T. Montelione, D. Zheng, Y.J. Huang, K.C. Gunsalus, T. Szyperski, Nat. Struct. Biol. 7 (suppl.) (2000) 982–985.

[22] H.N.B. Moseley, D. Monleon, G.T. Montelione, Methods Enzymol. 339 (2001) 91–108.

[23] H.N.B. Moseley, G.T. Montelione, Curr. Opin. Struct. Biol. 9 (1999) 635–642.

[24] V.Y. Orekhov, I.V. Ibraghimov, M. Billeter, J. Biomol. NMR 20 (2001) 49–60.

[25] M. Ottiger, F. Delaglio, A. Bax, J. Magn. Reson. 131 (1998) 373–378.

[26] J.H. Prestegard, H. Valafar, J. Glushka, F. Tian, Biochemistry 40 (2001) 8677–8685.

[27] S. Russell, P. Norvig, Artificial Intelligence, Prentice Hall, Saddle River, NJ, 1995.

[28] J. Simorre, B. Brutscher, M.S. Caffrey, D. Marion, J. Biomol. NMR 4 (1994) 325–333.

[29] R.R. Sokal, F.J. Rohlf, Biometry, W.H. Freeman and Company, New York, 1995..

[30] T. Szyperski, G. Wider, J.H. Bushweller, K. Wüthrich, J. Am. Chem. Soc. 115 (1993) 9307–9308.

[31] T. Szyperski, M. Pellecchia, K. Wüthrich, J. Magn. Reson. B 105 (1994) 188–191.

[32] T. Szyperski, D. Braun, C. Fernandez, C. Bartels, K. Wüthrich, J. Magn. Reson. B 108 (1995) 197–203.

[33] T. Szyperski, D. Braun, B. Banecki, K. Wüthrich, J. Am. Chem. Soc. 118 (1996) 8146–8147.

[34] T. Szyperski, B. Banecki, D. Braun, R.W. Glaser, J. Biomol. NMR 11 (1998) 387–405.

[35] T. Szyperski, D.C. Yeh, D.K. Sukumaran, H.N.B. Moseley, G.T. Montelione, Proc. Natl. Acad. Sci. USA 99 (2002) 8009–8014.

[36] D.E. Zimmerman, C.A. Kulikowski, Y. Huang, W. Feng, M. Tashiro, S. Shimotakahara, C. Chien, R. Powers, G.T. Montelione, J. Mol. Biol. 269 (1997) 592–610.