Journal of Biomolecular NMR **28:** 341–355, 2004.
KLUWER/ESCOM

Assignment validation software suite for the evaluation and presentation of protein resonance assignment data

Hunter N.B. Moseley[a], Gurmukh Sahota[a] & Gaetano T. Montelione[a,b,*]
[a]*Department of Molecular Biology and Biochemistry, Center for Advanced Biotechnology and Medicine, Rutgers University, Piscataway, NJ 08854, U.S.A.;* [b]*Department of Biochemistry and Molecular Biology, Robert Wood Johnson Medical School, Piscataway, NJ 08854, U.S.A.*

Received 9 July 2003; Accepted 10 November 2003

Key words: AutoAssign, BioMagResBank, structural proteomics

Abstract

We present a set of utilities and graphical user interface (GUI) tools for evaluating the quality of protein resonance assignments. The Assignment Validation Software (AVS) suite, together with new GUI features in the AutoAssign software package, provides a set of reports and graphs for validating protein resonance assignment data before its use in structure analysis and/or submission to the BioMagResBank (BMRB). Input includes a listing of resonance assignments and a summary of sequential connectivity data (i.e. triple resonance, NOE, or other data) used in deriving the assignments. These tools are useful for evaluating the accuracy of protein resonance assignments determined by either automated or manual methods.

Abbreviations: AVS, Assignment Validation Software suite of tools for validating protein resonance assignments; BMRB, BioMagResBank; CMap file, Connectivity Map file in a specific format; CMI, Connectivity Map Image; CMI Editor, tool for editing Connectivity Map Images; CSI, Chemical Shift Index; NESG, Northeast Structural Genomics Consortium; PRTL, Possible Residue Type List; SRO, Shift Relative Order validation analysis; 3D, three-dimensional.

Introduction

Recent advances in sample preparation, hardware for data collection, and software for automated analysis provide a significant reduction in the time required to generate sequence-specific resonance assignments and 3D structures of proteins using solution NMR methods (Moseley and Montelione, 1999; Montelione et al., 2000; Herrmann et al., 2002; Huang et al., 2003). In particular, recently established international efforts in structural genomics include several research consortia involved in 'high throughput' analysis of protein resonance assignments and solution NMR structures (Brenner, 2001). Most, if not all, of the resonance assignment data are deposited into the public domain BioMagResBank (BMRB) (Seavey et al., 1991) database of chemical shift and other NMR data for biomolecules. These resonance assignments provide the basis for structural and functional studies of proteins, and the resulting chemical shift database is a tremendously valuable long-term resource for the scientific community.

As part of the deposition process, the BMRB carries out a set of loose validations checking chemical shift assignments against expected values (their mean and standard deviations) derived from the BMRB archive. This consistency check reports resonance assignments that are five standard deviations outside their expected values (Eldon Ulrich, pers. commun.). This is a critical check aimed at preventing egregious assignment errors from entering the BMRB archive. In addition, the quality of a particular set of protein resonance assignments should be evaluated with respect to experimental data from which they are derived in

*To whom correspondence should be addressed. E-mail: guy@cabm.rutgers.edu

addition to comparison to expected values. Usually, this evaluation occurs hand-in-hand with the assignment process as a consistency check. The assignment process, by its nature, is exploratory as multiple possible assignments are considered during the directed search for a solution. Accordingly, this type of consistency check for an exploratory process must be less stringent, or loose, so that multiple possible assignments are detected and considered especially when resolving quality issues in the spectral data. However, detection of errors in complete or near-complete assignments requires a stricter evaluation. It is therefore critical, particularly in efforts which process large volumes of NMR assignment data, to have objective tools that can broadly evaluate the assignments against both the spectral data from which they are derived and the database of chemical shift distributions of specific protein atom types.

Recently, methods using chemical shifts calculated from a structure to validate chemical shift assignments (Oldfield, 1995; Wishart et al., 1997; Wishart and Case, 2001; Xu and Case, 2001) have been described. However, such methods are only useful once a three-dimensional (3D) structure is determined. As part of an effort to streamline the process of analyzing and archiving chemical shift assignments, we have developed a set of computer utilities for rigorously evaluating and validating a set of protein resonance assignments before submission to the BMRB and/or use in subsequent structure and/or functional analysis, without the need of a 3D structure. These tools, referred to as the Assignment Validation Software (AVS) suite, are summarized in Tables 1 and 2. They serve the purpose of providing strict consistency checks for detecting possible errors and identifying 'suspicious' assignments that deserve closer scrutiny prior to NOESY spectral analysis and 3D structure generation.

The AVS suite includes both new software tools and extensions of the GUI component of the AutoAssign software package (Zimmerman et al., 1997; Moseley et al., 2001). They are designed to perform the following tasks: (i) Statistical evaluation of individual chemical shifts and their associated amino-acid spin system classification against the database of protein chemical shift data, (ii) evaluation of the quality of information used to create segments of linked spin systems in the assignment process, and the uniqueness of their mapping into the protein amino acid sequence, and (iii) visual representation of assignment completeness and consistency with other spectral data not used in the assignment process but useful as additional validation of the assignment results.

Algorithms and implementation

The statistical evaluation of individual chemical shifts and their amino-acid classification in parts of the AVS suite involves comparison of the subject chemical shifts with values expected based on distributions of atom-type specific chemical shift values reported in the BMRB archive. These expected values and their variances come directly from the BMRB website (http://www.bmrb.wisc.edu/). These statistical measures exclude proteins with chemical shifts outside eight standard deviations from the raw mean or those that contain chemical shifts for at least one carbon bound proton greater than 10 ppm or less than -2.5 ppm, so as to eliminate from the calculations entries from paramagnetic proteins, proteins with aromatic prosthetic groups, or those containing unusual chemical shift referencing (http://www.bmrb.wisc.edu/ref_info/statsel.htm, 2002). Even with this culling, some remaining entries have assignment errors and slight referencing errors, which may result in some small bias of mean and standard deviation values; however, considering that these chemical shift data have significant deviations from normality due to secondary structure effects, these small differences in mean and standard deviations do not significantly affect the reliability of the validation algorithms that use these values. We refer to this set of atom-type specific chemical shift distributions, together with the corresponding atom-type specific mean, standard deviation, and min-max ranges, as the 'BMRB Expected Set'. The AVS suite can be updated with newer versions of the 'BMRB Expected Set'; however, the BMRB has matured to a point where the means and standard deviations of the 'BMRB Expected Set' do not significantly change from year to year for most resonance types.

Chemical shift assignment validation

The first program used in the AVS suite is **validate_assignments.pl** (Table 2), written in the Perl programming language. **Validate_assignments.pl** uses the 'BMRB Expected Set' to evaluate the chemical shift assignments in a BMRB entry. The program reads chemical shift values from a file in NMR-Star format, and calculates the chi-square probability

Table 1. Perl programs in assignment validation software (AVS) suite[a]

Perl program	Description
bmrb2cmap.pl	Converts BMRB format to CMap format.
compare_bmrb.pl	Compares two BMRB files and returns how well they match.
compare_cmap.pl	Compares two CMap files and returns how well they match.
create_assigned_peaklist.pl	Takes a BMRB file and assigns the spectral peaklists.
jval2bmrb.pl	Converts AutoStructure input $^3J(H^N\text{-}H^\alpha)$ values into BMRB coupling constants save frame.
missing_shifts.pl	Reports the missing Atom Types in a given BMRB file.
pks2bmrb.pl	Converts AutoAssign peak list into BMRB spectral peak list save frame.
sec2bmrb.pl	Converts AutoStructure output .sec file into BMRB secondary_structure save frame.
typing_degeneracy.pl	Calculates the mapping degeneracy for a given sequence.
validate_assignments.pl	Validates assignments in BMRB format.

[a]All components of the AVS suite are available at http://www-nmr.cabm.rutgers.edu/

Table 2. Summary of input data for key AVS processes

Process	Input type	Validation	Source
validate_assignments.pl	BMRB file with chemical shifts	chemical shift, spin system type	AVS suite
typing_degeneracy.pl	CMap file with amino acid sequence, chemical shift list, connectivity information	segment mapping	AVS suite
CMI Editor	CMap file	visualization	AutoAssign[a]

[a]Zimmerman et al. (1997); Moseley et al. (2001).

for each ^{1}H, ^{13}C and ^{15}N chemical shift assignment using the 'BMRB Expected Set'. This is similar to the validations that the BMRB performs on BMRB entries during the submission process; however, the **validate_assignments.pl** evaluation is posed as a stricter consistency check than the default analysis carried out on submission of data to the BMRB. **Validate_assignments.pl** reports assignments as 'suspicious' if the chi-square probability for the chemical shift is lower than a defined threshold (default is $p < 0.001$). The program reports assignments as 'potentially misassigned' if the chemical shift value falls outside the min-max range seen in the 'BMRB Expected Set' for that type of chemical shift. Furthermore, the program flags chemical shift assignments associated with atom types not included in its standard library, and those that have duplicate entries in the chemical shift list. This cleans up common clerical errors that arise in manual and semiautomated assignment methods.

Next, **validate_assignments.pl** performs an 'amino acid type' analysis (i.e., a validation of the residue-type assigned to the subject spin system). The assigned aliphatic carbon chemical shifts for each residue are evaluated using Bayesian methodology (Zimmerman et al., 1997) and chi-square probabilities (Equation 1) to calculate Bayesian class posterior probabilities (Duda and Hart, 1973) (Equation 2).

$$P(V_c|R) = P\left(\chi^2_N \geq \sum_j^N \left(\frac{V_{Cj} - \mu_{Rj}}{\sigma_{Rj}}\right)^2\right), \quad (1)$$

where V_C is the vector of carbon chemical shifts; R, residue type; N, number of carbon chemical shifts in V_c comparable to residue type R; μ_{Rj}, expected (mean) value for the jth comparable carbon chemical shift in V_C; σ_{Rj}, standard deviation for the jth comparable carbon chemical shift in V_C; χ^2_N, the chi squared value of the probability density function for a chi square distribution of N degrees of freedom that is greater than or equal to the test statistic; $P(\chi^2_N \geq$ test statistic), the area under the χ^2_N probability density function for values of χ^2_N greater than or equal to the test statistic.

$$P(R|V_C) = \frac{P(V_C|R)P(R)}{\sum_i P(V_C|Ri)P(Ri)}, \quad (2)$$

where V_C is the vector of carbon chemical shifts; R, residue type; $P(V_C|R)$, chi square probability of carbon chemical shifts given the residue type; $P(R)$, prior probability of given residue type, taken here, to be 1.0; i, index of residue type.

Original methods for carbon chemical shift based amino acid type analysis compared Cα and Cβ chemical shifts to random coil values and treated all amino acids as having a common Cα and Cβ chemical shift distribution (Grzesiek and Bax, 1992). This was appropriate considering the limited number of well-referenced protein chemical shift assignments available at that time. In contrast, the $P(V_c|R)$ chi square probability compares all available aliphatic carbon chemical shifts to their expected values and individual variances for folded proteins (BMRB Expected Set), thus handling spin systems with incomplete carbon chemical shift assignments. In addition, this is an improvement over the implementation of a similar Bayesian probability in AutoAssign (Zimmerman et al., 1997), which only uses Cα and Cβ chemical shifts.

Validate_assignments.pl performs the analysis on either partial or complete chemical shift lists. The input can be optionally supplemented with unassigned (H)CC(CO)NH TOCSY (Logan et al., 1992; Montelione et al., 1992) or HN(COCACB)CG (Constantine et al., 1997) peak lists, improving the residue-type validation using additional side chain carbon chemical shift information available from these data.

Each potential residue-type for a particular residue is rank ordered by its Bayesian class posterior probability, $P(R|V_c)$, and the top candidates whose probabilities sum above a threshold (default is $\Sigma_i P(R|V_c)_i > 0.999$) are identified. This generally corresponds to 6–10 possible residue types for a given amino acid spin system. The program reports residue spin system assignments as 'suspicious' when the reported residue type is not included in this list of residue-types with highest Bayesian class posterior probabilities. The program reports residue spin system assignments as 'potentially mistyped' when the reported residue type is not included in the list of top conditional probabilities even when specific carbon assignments are relaxed, allowing all carbon chemical shifts of a residue to permute through all carbon shift assignment possibilities for that residue.

Finally, **validate_assignments.pl** compares the relative order of assigned chemical shifts in the same residue against 'shift relative order' (SRO) rules. An analysis of the BMRB archive revealed this set of SRO rules and the degree of consistency observed in the BMRB archive for each rule. For example, 99.86% of the time the tyrosine Cα resonance is downfield of the tyrosine Cβ resonance in 1450 tyrosine residues from the BMRB archive. The consistency for the rule that tyrosine Cα chemical shift value is greater than tyrosine Cβ chemical shift value is 0.9986. The program reports broken SRO rules of a given consistency or higher (default consistency is 0.99).

As described above, the **validate_assignments.pl** Perl program performs 'atom-type', 'residue-type', and 'SRO' consistency checks. Together, these three checks comprise an evaluation based solely on the expected values obtained from statistical distributions of chemical shift data reported in the BMRB. In addition, the program has a host of options for customizing the stringency of these checks, and for handling experiment- and sample-specific features of the chemical shift list. These include options for handling chemical shift assignments arising from spectra collected on full and partial uniformly deuterated samples.

For perdeuterated protein samples, **validate_assignments.pl** adjusts the ^{13}C (excluding C′) and ^{15}N chemical shifts in the 'BMRB Expected Set' according to the specified amount of deuteration and then performs the three consistency checks described above. These adjustments (Table 3) are empirically derived from a combination of sources including comparisons of fully deuterated and nondeurated proteins for average deviation of residue-type-specific Cα chemical shifts (Venters et al., 1996), estimates of deuterium isotope effects on nitrogen (Hansen, 1988; Gardner and Kay, 1998), and analysis of deuterated small molecules with similar chemical structure to the sidechains of the 20 amino acids (Forsyth, 1984; Hansen, 1988). Individual one-bond, two-bond, and three-bond deuterium isotope effects were summed to reach full deuterium correction values for each carbon and nitrogen resonance type. These deuteration correction values are then subtracted from the values of the 'BMRB Expected Set'. For partially deuterated samples, the full deuteration correction values are multiplied by the fraction of deuteration (assuming uniform partial deuteration) before being subtracted from the 'BMRB Expected Set' values.

Table 3. Deuterium isotope shift corrections

ResonType	Correction	ResonType	Correction	ResonType	Correction	ResonType	Correction
A-CA	0.752	H-CA	0.632	M-CA	0.716	T-CA	0.638
A-CB	0.932	H-CB	0.832	M-CB	0.917	T-CB	0.867
A-N	0.34	H-CG	0.365	M-CG	0.995	T-CG	0.950
		H-CD2	0.545	M-CE	0.855	T-N	0.22
C-CA	0.632	H-CE1	0.475	M-N	0.28		
C-CB	0.837	H-N	0.28			V-CA	0.764
C-N	0.28	H-ND1	0.34	N-CA	0.632	V-CB	1.057
		H-NE2	0.32	N-CB	0.757	V-CG1	0.950
D-CA	0.632			N-N	0.28	V-CG2	0.950
D-CB	0.690	I-CA	0.722			V-N	0.22
D-N	0.28	I-CB	1.092	P-CA	0.716		
		I-CG1	1.130	P-CB	1.027	W-CA	0.632
E-CA	0.716	I-CG2	1.020	P-CG	1.158	W-CB	0.832
E-CB	0.937	I-CD1	1.090	P-CD	0.952	W-CG	0.365
E-CG	0.830	I-N	0.22	P-N	0.72	W-CD1	0.350
E-N	0.28					W-CD2	0.213
		K-CA	0.716	Q-CA	0.716	W-CE2	0.167
F-CA	0.632	K-CB	1.067	Q-CB	0.951	W-CE3	0.429
F-CB	0.919	K-CG	1.225	Q-CG	0.915	W-CZ2	0.429
F-CG	0.274	K-CD	1.190	Q-N	0.28	W-CZ3	0.511
F-CD1	0.429	K-CE	0.990			W-CH2	0.511
F-CD2	0.429	K-N	0.28	R-CA	0.716	W-N	0.28
F-CE1	0.500	K-NZ	0.44	R-CB	1.067	W-NE1	0.22
F-CE2	0.500			R-CG	1.155		
F-CZ	0.496	L-CA	0.674	R-CD	0.990	Y-CA	0.632
F-N	0.28	L-CB	1.067	R-CZ	0.070	Y-CB	0.919
		L-CG	1.215	R-N	0.28	Y-CG	0.299
G-CA	0.784	L-CD1	1.090	R-NE	0.44	Y-CD1	0.418
G-N	0.32	L-CD2	1.090			Y-CD2	0.418
		L-N	0.28	S-CA	0.632	Y-CE1	0.390
				S-CB	0.837	Y-CE2	0.390
				S-N	0.28	Y-CZ	0.220
						Y-N	0.28

Figure 1 shows a sample of the report generated by the **validate_assignments.pl** program. The report provides a brief validation summary for each residue, together with an overall summary of the most severe potential errors of atom-type assignments, residue-type assignments, and SRO's. There is also a report summarizing the total number of detected 'potential errors' and 'suspicions' for each category, and a consolidated list of all 'potential errors' and 'suspicions'.

Segment mapping degeneracy analysis

The AVS suite also performs a *segment mapping degeneracy analysis,* evaluating the quality of information used to define and assign 'segments of linked spin systems'. This analysis finds 'segments of linked spin systems' (Zimmerman et al., 1997) indicated by triple-resonance and/or NOESY NMR data and assesses the quality of data supporting the links between the dipeptide spin systems and the uniqueness of their mapping to the amino acid sequence. The **typing_degeneracy.pl** Perl program (Table 2) performs this analysis in four stages, using as input (i) the complete amino acid sequence of the subject protein, (ii) the chemical shift assignments, (iii) a list of intraresidue and sequential spin system connectivities generated either by manual analysis of triple resonance NMR data or by automated analysis with

```
M1  Overall: Uncertain   Residue Type: Uncertain   SRO: Uncertain   C Shifts: Uncertain   H Shifts: Uncertain

A2  Overall: Consistent   Residue Type: Consistent   SRO: Consistent   C Shifts: Consistent   H Shifts: Consistent

    Possible Residue Type List (PRTL)>>  A 0.8541  M 0.1374  L 0.0064  V 0.0009  I 0.0007  T 0.0001

    C Shift Assignments>>  C :: 179.3  CA :: 51.5  CB :: 19.4

    H Shift Assignments>>  H N:: 8.57  HA :: 4.04  HB :: 1.46

K3  Overall: Misassigned   Residue Type: Suspicious   SRO: Consistent   C Shifts: Misassigned   H Shifts: Consistent

    Possible Residue Type List (PRTL)>>  M 0.7431  Q 0.184  E 0.0552  Y 0.0099  R 0.0055  F 0.0011  I 0.0009

    C Shift Assignments>>  C :: 176.4   CA :: 56.6   CB :: 32.7   CG :: 32.7(M)
    Expected C Shift Values >>   C :: 176.46(2.05)  CA :: 56.84(2.25)  CB :: 32.83(1.88)  CG :: 24.91(1.31)  CD :: 28.78(1.39)  CE :: 41.78(0.98)

    H Shift Assignments>>  H N:: 8.58  HA :: 4.14  HB2 :: 1.82  HB3 :: 1.82  HG2 :: 1.51  HG3 :: 1.51
...
F5  Overall: Consistent   Residue Type: Consistent   SRO: Consistent   C Shifts: Consistent   H Shifts: Consistent

    Possible Residue Type List (PRTL)>>  Y 0.9441  F 0.0557

    HN Overlap>>  M35  D63  R90

    C Shift Assignments>>  C :: 176.1  CA :: 57.5  CB :: 39.7  CD1 :: 128.3  CD2 :: 128.3  CE1 :: 132.06
                           CE2 :: 132.06

    H Shift Assignments>>  HN :: 8.47  HA :: 4.58  HB2 :: 3.12  HB3 :: 2.93  HD1 :: 7.04  HD2 :: 7.04
                           HE1 :: 7.32  HE2 :: 7.32
...
Error Summary:

K3  Residue Type: Suspicious
K3  CG = 32.7(M),          Expected = 24.91, StD = 1.31, ChiSquare = 2.7383e-09
P8  HD2 = 0.82(M),         Expected =  3.64,  StD = 0.36, ChiSquare = 4.7510e-15
P8  HD3 = 0.82(M),         Expected =  3.63,  StD = 0.40, ChiSquare = 2.1407e-12
Q9  Residue Type: Mistyped
Q9  CD = 29.1(M),          Expected = 179.68, StD = 1.17, ChiSquare = 0.0000e+00
Q9  HG2 = 3.31(S),         Expected =  2.32,   StD = 0.29, ChiSquare = 6.4065e-04
Q9  HG3 = 3.31(S),         Expected =  2.32,   StD = 0.29, ChiSquare = 6.4065e-04
Q9  SRO Rule Break>>  CD > CA : 1
Q9  SRO Rule Break>>  CD > CB : 1
Q9  SRO Rule Break>>  CD > CG : 1
V11 CB = 41.1(M),          Expected = 32.66,  StD = 1.91, ChiSquare = 9.9228e-06
K17 Residue Type: Suspicious
K17 CD = 42.7(M),          Expected = 28.78,  StD = 1.39, ChiSquare = 1.3177e-23
...
```

Figure 1. Validation report generated by the **validation_assignments.pl** analysis of an intermediate chemical shift list for *E. coli* ribosomal binding factor A (RbfA). The report begins with the chemical shift validation results for RbfA assignments broken down by residue. The first line shows the overall status for each residue and then the status for each of the five categories tested. Possible values of status include consistent, suspicious, (possibly) misassigned, and uncertain. Next comes the Possible Residue Type List (PRTL) with the list of possible amino acid types associated with the corresponding ^{13}C chemical shift data, based on Bayesian statistical analysis (Equation 2). The next line, HN Overlap, provides a list of 'overlapped' spin systems with similar H^N-N chemical shift values, if they exist. The next lines show the ^{13}C and ^{1}H chemical shift assignments and any errors detected. Errors are indicated with an (M) for 'possibly misassigned', (S) for 'suspicious', (D) for 'duplicate entry', or (U) for 'unknown' chemical shift type. When assignments are characterized as 'suspicious' or 'possibly misassigned', the expected ^{13}C and ^{1}H chemical shifts values and corresponding standard deviations (in parenthesis) for that residue type are also shown. The next line shows Shift Relative Order (SRO) errors if they exist (not shown in this example). An Error Summary is also provided at the end of the validation report, listing all possible errors encountered for each residue.

programs like AutoAssign (Moseley et al., 2001), and/or (iv) a set of sequential NOEs generated manually or automatically with programs like AutoStructure (Huang, 2001; Huang et al., 2003) or CYANA (Hermann et al., 2002). **Typing_degeneracy.pl** reads this information in the form of a Connectivity Map File (CMap file), a standard file format used by AutoAssign and AVS.

In the first stage of the *segment mapping degeneracy analysis*, **typing_degeneracy.pl** scans the spin system connectivity data of each sequential pair of assigned spin systems for evidence of linkage. This comes in the form of matching intraresidue and sequential connectivity data and/or local NOESY constraints that span the sequential pair of assigned spin systems. The program represents the interresidue connectivity data as a 'linkage strength', a count of each unambiguously matched sequential connection and spanning local NOESY constraint for the sequential spin system pair. The program then groups sets of sequentially connected residue pairs that have a linkage strength equal to or above a threshold (default is 2) into a 'linked segment of spin systems' (a GS segment in AutoAssign terminology). Spin systems that are at the ends of segments and have a 'linkage strength' less than the threshold are marked 'possibly misassigned'.

In the next stage of the *segment degeneracy analysis*, **typing_degeneracy.pl** uses the assigned aliphatic carbon chemical shifts of each spin system to calculate the residue-type Bayesian class posterior probability for each assigned spin system (Equation 2). The program then uses these residue-type probabilities to calculate a *segment mapping likelihood* for each linked segment to each location in the sequence (Equation 3). The amino acid residue type analysis carried out by the typing_degeneracy.pl program is different from that described above for the amino acid residue type analysis of the validate_assignments.pl program. Both programs perform Bayesian-based amino acid type analysis but for different posed questions. In validate_assignments.pl, the algorithm evaluates the set of amino acid types that fit the given ^{13}C chemical shift data for a particular spin system. In typing_degeneracy.pl, a different algorithm is used to evaluate how uniquely a set of linked ^{13}C chemical shifts map to a stretch of amino acid residue types. This latter evaluation is different, and actually, more strict, than the former.

$$\text{Segment Mapping Likelyhood} = \prod_{i=1}^{N} \frac{P(S_{k+i-1}|V_{Ci})}{P(R_{Top_i}|V_{Ci})}, \tag{3}$$

where S_j is the amino acid type at sequence site j; $P(S_{k+i-1}|V_{Ci})$, probability of amino acid type at sequence position $k + i - 1$ given the vector V_{Ci} of carbon chemical shifts; R_{Topi}, amino acid type with the highest probability for the given vector V_{Ci} of carbon chemical shifts; $P(R_{Topi}|V_{Ci})$, highest amino acid type probability given the vector V_{Ci} of carbon chemical shifts.

The $P(R_{Topi}|V_{Ci})$ probability normalizes the *segment mapping likelihood* based on the discriminating power of the given chemical shifts V_{Ci}. This simplifies the interpretation of the *segment mapping likelihood* since smaller values strictly indicate the mapping to lower probable amino acid types and not the presence of spin systems lacking carbon chemical shifts. This differs from AutoAssign's implementation of a similar segment mapping score which has no normalization.

Next, the program performs a 'likelihood evaluation'. For each linked segment, the program compares its *reported segment mapping likelihood,* corresponding to the mapping indicated in the chemical shift assignment list being evaluated, to its *segment mapping likelihoods* for all other possible locations in the amino acid sequence. If the *segment mapping likelihood* for the site reported in the chemical shift list is at least X times larger than all other segment mapping likelihoods (default is $X = 1000$), then the mapping for this linked segment is deemed unique. All non-unique mappings are marked for later analysis. The program MAPPER (Güntert et al., 2000) carries out a similar segment mapping analysis but for the purpose of semi-automated backbone resonance assignments of spin system segments.

In the next stage of the *segment degeneracy analysis*, the **typing_degeneracy.pl** program analyzes the amino acid sequence for inherent amino acid type degeneracies due to the presence of *similar sequence segments.* Such degeneracy can lead to 'swapping' of linked spin system segments between different potential mappings into the amino acid sequence. We have derived a *similarity matrix* characterizing the inherent ambiguity in spin system identification based on Cα/Cβ (Table 4) and Cα/Cβ/Cγ (Table 5) chemical shift information. We calculated all residue-type probabilities using Equation 2 with chemical shift data for 40,592 residues in the BMRB archive. Data was included only for non-glycine residues for which both

Table 4. Similarity matrix describing residue-type similarity based on Cα/Cβ chemical shift data

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	1.0	0.5203	0.0000	0.0080	0.0000	0.2184	0.1036	0.0000	0.0010	0.0000	0.0349	0.0000	0.0000	0.0246	0.0122	0.0000	0.0000	0.0000	0.0141	0.0000
C	0.0085	1.0	0.4048	0.6429	0.4592	0.0289	0.6718	0.4371	0.7415	0.3895	0.7874	0.4167	0.3452	0.6088	0.6548	0.0000	0.0000	0.6140	0.5000	0.5170
D	0.0000	0.5958	1.0	0.0226	0.9746	0.0472	0.1283	0.9770	0.3211	0.9891	0.8201	0.9883	0.0004	0.0117	0.0528	0.0000	0.0000	0.2969	0.0444	0.9447
E	0.0040	0.5747	0.0401	1.0	0.0855	0.0023	0.8743	0.0546	0.9872	0.0155	0.9197	0.1520	0.4339	0.9717	0.9704	0.0000	0.0000	0.7286	0.6303	0.1635
F	0.0000	0.6016	0.9724	0.0631	1.0	0.0033	0.2058	0.9744	0.3610	0.9882	0.7692	0.8494	0.0151	0.0250	0.0881	0.0000	0.0000	0.5233	0.0664	0.9546
G	0.7563	0.5768	0.3280	0.0295	0.1655	1.0	0.2524	0.0036	0.0732	0.2289	0.2378	0.4786	0.0000	0.1077	0.1926	0.0083	0.0009	0.0012	0.0949	0.1232
H	0.0260	0.6236	0.0488	0.9892	0.1345	0.0130	1.0	0.1150	0.9610	0.0228	0.8948	0.1974	0.2549	0.9729	0.9707	0.0000	0.0000	0.4967	0.7321	0.2245
I	0.0000	0.5685	0.6741	0.0635	0.9631	0.0000	0.2283	1.0	0.3169	0.8219	0.6641	0.5637	0.1388	0.0298	0.0985	0.0000	0.0000	0.6741	0.0682	0.9365
K	0.0006	0.5652	0.1931	0.9815	0.6678	0.0058	0.8648	0.5368	1.0	0.1132	0.9284	0.5585	0.4079	0.9434	0.9552	0.0000	0.0000	0.9341	0.6662	0.7964
L	0.0000	0.5755	0.9855	0.0053	0.9760	0.0246	0.0324	0.9789	0.0828	1.0	0.5340	0.9766	0.0000	0.0030	0.0116	0.0003	0.0000	0.0864	0.0098	0.9418
M	0.0067	0.6157	0.2461	0.9506	0.5629	0.0124	0.8573	0.4427	0.9809	0.1798	1.0	0.6180	0.2876	0.9157	0.9461	0.0000	0.0000	0.8753	0.6292	0.7202
N	0.0000	0.5988	0.9825	0.2765	0.9740	0.1270	0.6050	0.9345	0.8177	0.9532	0.8832	1.0	0.0000	0.1343	0.4047	0.0000	0.0000	0.7788	0.3426	0.9458
P	0.0000	0.6276	0.0026	0.7364	0.5285	0.0000	0.8874	0.8932	0.9757	0.0109	0.8784	0.0045	1.0	0.3263	0.9162	0.0000	0.0000	0.9936	0.7012	0.8445
Q	0.0173	0.6235	0.0155	0.9969	0.0347	0.0074	0.9022	0.0186	0.9697	0.0050	0.9257	0.1084	0.2533	1.0	0.9802	0.0000	0.0000	0.3932	0.6687	0.0768
R	0.0031	0.6176	0.0621	0.9917	0.1231	0.0104	0.8946	0.0866	0.9854	0.0167	0.9280	0.2081	0.3667	0.9718	1.0	0.0000	0.0000	0.7590	0.6792	0.2572
S	0.0000	0.0159	0.0000	0.0000	0.0000	0.0012	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0	0.3059	0.0000	0.0000	0.0000
T	0.0000	0.0004	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2327	1.0	0.0000	0.0000	0.0000
V	0.0000	0.5920	0.1347	0.6843	0.6562	0.0000	0.8201	0.9252	0.9157	0.1960	0.7606	0.2073	0.9007	0.4131	0.8471	0.0000	0.0000	1.0	0.6504	0.8482
W	0.0065	0.6450	0.0390	0.9957	0.1104	0.0065	0.9199	0.0866	0.9784	0.0217	0.9351	0.1667	0.4242	0.9827	0.9762	0.0000	0.0000	0.7424	1.0	0.2273
Y	0.0000	0.5936	0.9679	0.1108	0.9671	0.0072	0.3631	0.9711	0.5100	0.9767	0.8137	0.8779	0.0426	0.0562	0.1518	0.0000	0.0000	0.6305	0.1309	1.0

Cα and Cβ chemical shifts were reported or glycine Cα shifts. Residues with significant chemical shift outliers were excluded. Next, we determined the fraction of times that residues of a given residue type had a residue-type probability for another specific residue type included in the top list of residue-type probabilities summing to a cumulative probability total $\Sigma_i P(R|V_c)_i > 0.9999$. This generates the elements in the Cα/Cβ similarity matrix (Table 4) excluding diagonal elements which were set to 1. We also generated a Cα/Cβ/Cγ similarity matrix (Table 5), using the same procedure and 32 216 residues with appropriate chemical shift data. Using *similarity matrices*, **typing_degeneracy.pl** calculates the *segment similarity scores* (Equation 4) between all pairs of *sequence segments* (an SS segment in AutoAssign terminology) of a given length in the amino acid sequence.

$$\text{Segment Similarity}\,(j,k,L) \;=\; \prod_{i=0}^{L-1} \mathbf{M}_{S_{j+i},S_{k+i}}, \quad (4)$$

where **M** is the amino acid type similarity matrix; $\mathbf{M}_{x,y}$, matrix element at row x and column y; S_x, residue type at position x in the sequence; L, length of residue segment to compare; i, index over segment length L; j,k, indices corresponding to the first positions of the sequence segments being compared.

By default, **typing_degeneracy.pl** uses the *similarity matrix* **M** based on Cα/Cβ chemical shift data (Table 4).

The program looks for *segment similarity scores* above a certain threshold (default value is 0.5). When it finds such *similar sequence segments*, it marks them as 'similar' and sets the *degeneracy length* parameter assigned to each residue in that particular segment to the segments' length. The program first searches for *similar sequence segments* that are 2 residues long. Next it increments this length (L) and searches for *similar sequence segments* at length L. This process is iterated until no more *similar sequence segments* are found above the threshold. Searching for *similar sequence segments* in this manner determines the (maximum) *degeneracy length* associated with each residue in the protein sequence. Next, the program performs a 'similarity evaluation': The program marks each *linked spin system segment* completely mapped within a *similar sequence segment* for later analysis, because there exist alternate stretches of the amino acid sequence with similar C^{α}/C^{β} (or $C^{\alpha}/C^{\beta}/C^{\gamma}$) chemical shift profiles.

Finally, the program performs an analysis of *linked spin system segments* to identify those that are potentially incorrectly mapped into the protein sequence. The program starts by labeling each *linked spin system segment* marked during 'likelihood evaluations' and 'similarity evaluations' as 'possibly misassigned'. The program iteratively evaluates such marked *linked spin system segments* to see if their alternate mappings have been independently *verified.* A *verified* alternate mapping is a sequence segment assigned to a *linked spin system segment* that is not labeled as 'possibly misassigned'; i.e., one that is unambiguously assigned. The 'possibly misassigned' *linked spin system segments* for which all alternate mappings are so *verified* are relabeled as simply 'suspicious'. The program iterates until no additional 'possibly misassigned' *linked spin system segments* are relabeled as 'suspicious'. The remaining 'possibly misassigned' linked spin system segments, for which there exist alternate mappings that are not verified, thus remain labeled as 'possibly misassigned', and require careful manual inspection.

After completing all three stages of the segment degeneracy analysis, the **typing_degeneracy.pl** Perl program generates a report in two parts: (i) a list of 'suspicious' or 'possibly misassigned' *linked spin system segments* and (ii) a list of *similar sequence segments.* An example from a section of this report is shown in Figure 2. The list of suspect *linked spin system segments* (left side of Figure 2) includes the *segment mapping likelihoods* (Equation 3) and their ratios for alternative mappings [Ratio = *segment mapping likelihood* (assigned)/*segment mapping likelihood* (alternative)].

These ratios of *segment mapping likelihoods* are measures of the mapping uniqueness of suspicous *linked spin system segments*; high values (>1000) indicate a high likelihood that the assigned mapping is unique. Lower values of these ratios indicate the presence of good alternative mappings of the linked amino acid spin systems into the amino acid sequence, and thus, a less unique mapping. The list of *similar sequence segments* (right side of Figure 2) includes segment similarity scores which indicate how similar the segments are based on spin system identification. Higher segment similarity scores represent more similar *sequence segments.* The statistics summarized in Figure 2 are valuable for validating the mapping of *linked spin system segments* into the amino acid sequence, and for identifying potential erroneous mappings.

Table 5. Similarity matrix describing residue-type similarity based on Cα/Cβ/Cγ chemical shift data

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	1.0	0.5203	0.0000	0.0079	0.0000	0.2184	0.1036	0.0000	0.0009	0.0000	0.0348	0.0000	0.0000	0.0246	0.0121	0.0000	0.0000	0.0000	0.0140	0.0000
C	0.0085	1.0	0.4047	0.6428	0.4591	0.0289	0.6717	0.4370	0.7414	0.3894	0.7874	0.4166	0.3452	0.6088	0.6547	0.0000	0.0000	0.6139	0.5000	0.5170
D	0.0000	0.6045	1.0	0.0235	0.9739	0.0474	0.1305	0.9764	0.3211	0.9890	0.8228	0.9882	0.0004	0.0121	0.0545	0.0000	0.0000	0.2993	0.0457	0.9450
E	0.0080	0.5969	0.0000	1.0	0.0000	0.0055	0.0000	0.0145	0.0015	0.0005	0.7688	0.0000	0.0005	0.9507	0.0105	0.0000	0.0000	0.0000	0.0000	0.0000
F	0.0000	0.6006	0.9721	0.0635	1.0	0.0033	0.2066	0.9741	0.3622	0.9880	0.7688	0.8503	0.0152	0.0251	0.0887	0.0000	0.0000	0.5251	0.0668	0.9543
G	0.7562	0.5767	0.3279	0.0294	0.1654	1.0	0.2523	0.0035	0.0732	0.2288	0.2377	0.4785	0.0000	0.1077	0.1925	0.0083	0.0008	0.0011	0.0949	0.1232
H	0.0239	0.6248	0.0490	0.9890	0.1352	0.0130	1.0	0.1155	0.9618	0.0229	0.8964	0.1984	0.2562	0.9727	0.9705	0.0000	0.0000	0.4983	0.7306	0.2257
I	0.0000	0.5773	0.0000	0.0000	0.0000	0.0006	0.0000	1.0	0.0106	0.0225	0.0033	0.0000	0.0026	0.0000	0.0019	0.0000	0.0000	0.2322	0.0000	0.0000
K	0.0016	0.5967	0.0000	0.0000	0.0000	0.0119	0.0000	0.4953	1.0	0.1392	0.0417	0.0000	0.4531	0.0005	0.9571	0.0000	0.0000	0.3631	0.0000	0.0000
L	0.0000	0.6041	0.0000	0.0000	0.0000	0.0358	0.0000	0.9603	0.1012	1.0	0.0114	0.0000	0.0000	0.0004	0.0076	0.0000	0.0000	0.0171	0.0000	0.0000
M	0.0111	0.6338	0.0000	0.4535	0.0000	0.0148	0.0000	0.3382	0.0483	0.0092	1.0	0.0000	0.0111	0.6914	0.6505	0.0000	0.0000	0.0000	0.0000	0.0000
N	0.0000	0.5987	0.9825	0.2765	0.9740	0.1269	0.6049	0.9345	0.8177	0.9531	0.8831	1.0	0.0000	0.1343	0.4046	0.0000	0.0000	0.7787	0.3425	0.9458
P	0.0000	0.6601	0.0000	0.0000	0.0000	0.0000	0.0000	0.8124	0.9585	0.0145	0.0300	0.0000	1.0	0.0010	0.7958	0.0000	0.0000	0.4103	0.0000	0.0000
Q	0.0250	0.6722	0.0000	0.9842	0.0000	0.0092	0.0000	0.0065	0.0018	0.0000	0.9210	0.0000	0.0018	1.0	0.0427	0.0000	0.0000	0.0000	0.0000	0.0000
R	0.0069	0.6342	0.0000	0.0086	0.0000	0.0156	0.0000	0.0651	0.9504	0.0130	0.6125	0.0000	0.3796	0.0165	1.0	0.0000	0.0000	0.0069	0.0000	0.0000
S	0.0000	0.0159	0.0000	0.0000	0.0000	0.0012	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0	0.3058	0.0000	0.0000	0.0000
T	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2508	1.0	0.0000	0.0000	0.0000
V	0.0000	0.6265	0.0000	0.0005	0.0000	0.0000	0.0000	0.9445	0.1649	0.0316	0.0000	0.0000	0.1047	0.0000	0.0046	0.0000	0.0000	1.0	0.0000	0.0000
W	0.0065	0.6441	0.0393	0.9956	0.1091	0.0065	0.9192	0.0851	0.9781	0.0218	0.9388	0.1637	0.4279	0.9825	0.9759	0.0000	0.0000	0.7445	1.0	0.2270
Y	0.0000	0.5911	0.9689	0.1095	0.9722	0.0065	0.3589	0.9713	0.5061	0.9779	0.8184	0.8765	0.0433	0.0539	0.1504	0.0000	0.0000	0.6287	0.1283	1.0

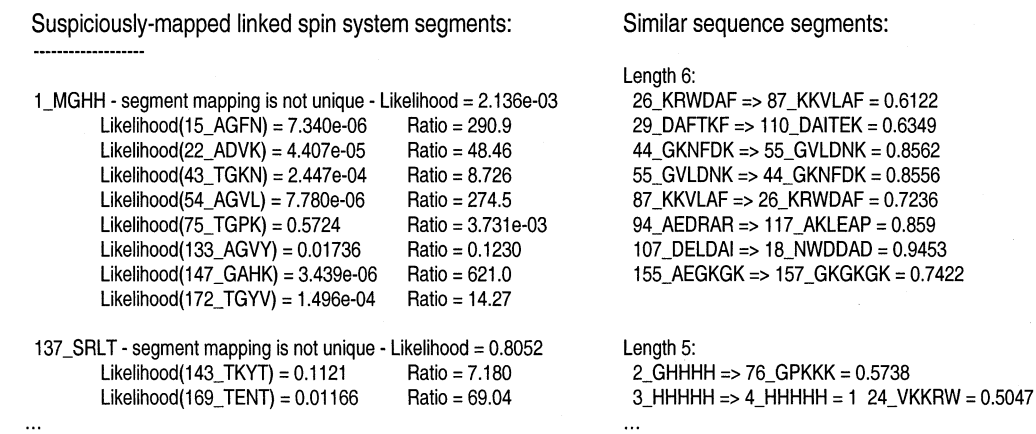

Figure 2. Validation report generated by **typing_degeneracy.pl** analysis of chemical shift list for *C. elegans* gene product ORF C32E8.3, NESG target WR33 (Monleon et al., 2003; BMRB accession number 5300). The left column shows suspect *linked spin system segments* in protein WR33. Each such segment of linked spin systems is listed along with the reason for its suspicion, its segment mapping likelihood (Equation 3), a list of other possible mapping locations, their segment mapping likelihoods, and the ratio of the segment mapping likelihoods for the original and alternate locations. The right column shows groups of *similar sequence segments* with inherent mapping degeneracy and their segment similarity scores (Equation 4).

Visual representation of assignment results

The final analysis of the AVS suite is a visual representation of assignment completeness and consistency for human inspection. For this purpose, we have developed the **Connectivity Map Image Editor** (**CMI Editor**) using part of the graphical user interface of the AutoAssign software package for visualizing sequential connectivity, segment degeneracy and other information that is stored in a CMap file. We use the CMap format instead of a NMRStar format because the CMap format contains additional parameters needed to visualize this resonance assignment data, and because it is easily manipulated by a user. The files are designed to provide easy editing with a simple text editor. Also, the CMap format handles partitioning of information into sections, allowing simple splitting and concatenation of CMap files.

Figures 3 and 4 show segments of images generated by the **CMI Editor.** The AutoAssign program generates an initial CMap file for visualizing the assignment completeness of its automated backbone assignments. This includes intraresidue H^N, N, C^α, C^β, C' and H^α chemical shift assignments and sequential C^α, C^β, C', and H^α chemical shift assignments. This information can be updated or edited based on manual analysis. The **typing_degeneracy.pl** Perl program can then analyze this CMap and generate a new CMap file with additional rows showing *linkage strengths* between each assigned spin system and the *degeneracy length* associated with each residue. Figure 3 shows the Connectivity Map Image (CMI) arising from the **typing_degeneracy.pl** analysis. *Linkage strength* shows the amount of connectivity data linking neighboring spin systems. The *degeneracy length* shows the longest *similar residue segment* that a residue belongs to. Shorter *degeneracy lengths* indicate *sequence segments* that are more unique. *Suspiciously-mapped linked spin system segments* are highlighted for easy visual recognition. Light blue (or gray) highlights on the degeneracy length row shows 'suspicious' segments for which alternative mappings have been independently assigned. Red (or black) highlights on the degeneracy length row indicate 'possibly misassigned' segments for which alternative mappings are not unambiguously assigned. Yellow highlights on the degeneracy length row indicate isolated spin systems at the end of mapped segments with low linkage strengths.

These CMap files may also be supplemented with additional manually-derived assignment information or created from fully manual analysis and displayed with the **CMI Editor** in AutoAssign. As shown in Figure 4, the CMI Editor can handle a variety of data including local NOE connectivities, secondary structure information, and scalar coupling data. User-defined data types, such as sequential connections established from residual dipolar couplings, heteronuclear NOE data, amide $^1H/^2H$ exchange data, etc, can be added and visualized as well, by simple editing of the CMap file. Figure 4 shows how the **CMI Editor** can be used to display secondary structure informa-

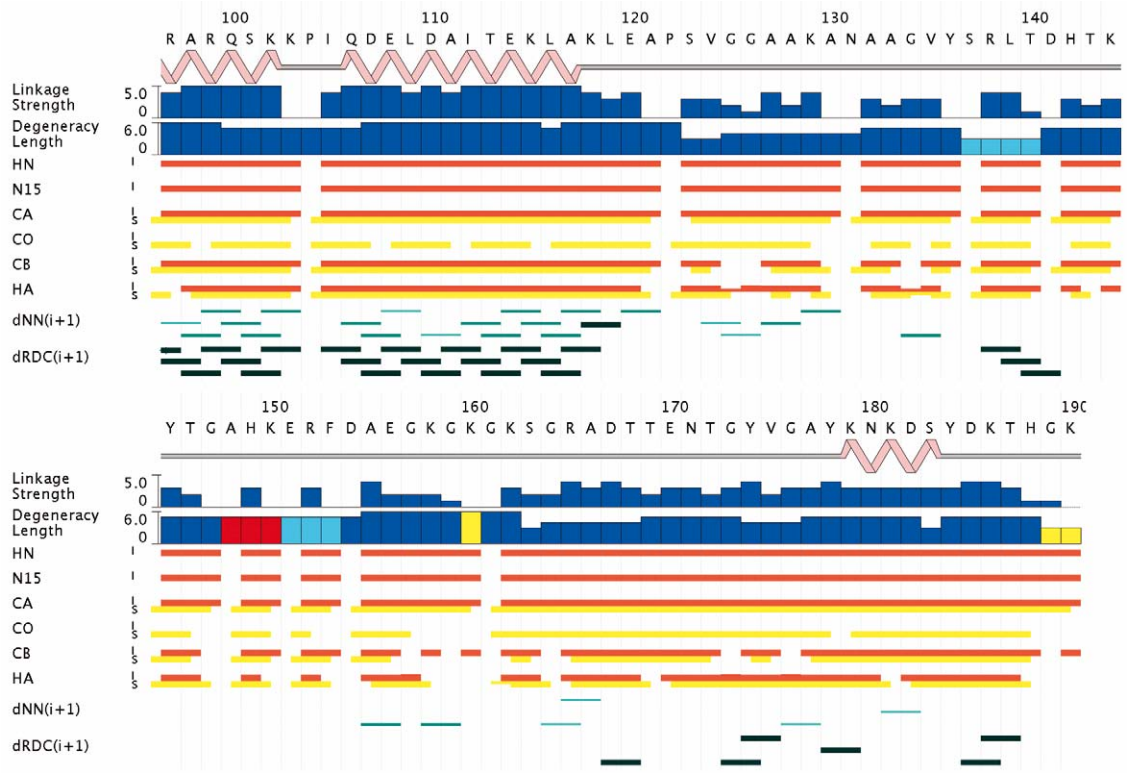

Figure 3. CMap Image from **typing_degeneracy.pl** Perl program analysis of chemical shift assignments for the C-terminal region of WR33 (BMRB accession number 5300). The first row is a partial sequence of protein WR33. The next row annotates the secondary structure. The next row shows the linkage strength between residue i and $i+1$ as a bar graph. Next is the maximum degeneracy length of each overlapping sequence segment for a residue represented in a bar graph. Red bars indicate assigned spin systems in a 'possibly misassigned spin system segment'. Light blue bars indicate assigned spin systems in a 'suspicious spin system segment'. Yellow bars indicate isolated spin systems at the end of mapped segments with low linkage strength. The next six rows are the chemical shift connectivity rows showing which resonances are assigned and whether the assignment came from intraresidue peaks, interresidue peaks, or both. The next row represents NOE connectivities used in making the assignments. The last row shows residual dipolar coupling connectivity data (Zweckstetter and Bax, 2001) also used in making these assignments (Monleon et al., 2003).

tion, intra and sequential triple-resonance data, local (intraresidue, sequential, and medium-range) NOE connectivities, scalar coupling data, and chemical shift index (CSI) information. The CSI analysis (Wishart and Sykes, 1994) for identifying secondary structure is done automatically by the **CMI Editor**. The CMap format allows the user to add arbitrary bar graph, symbol, and connectivity rows with user specified titles. The **CMI Editor** also provides facilities for manipulating the color, size, and order of all rows in the image. Once edited into its final form, the **CMI Editor** can generate high-resolution GIF formatted images suitable for publication.

Results

We developed and use the AVS suite in carrying out and validating resonance assignments of proteins being studied as part of the NIH Protein Structure Initiative in Structural Genomics. A particularly dramatic example involved the initial analysis of resonance assignments for the *E. coli* ribosomal binding protein A (RbfA). Using the AutoAssign program (Zimmerman et al., 1997; Moseley et al., 2001) and an initial set of problematic peak lists, a preliminary set of backbone resonance assignments were obtained which upon further investigation were found to contain significant numbers of misassignments (Table 6). Some 27 of the 105 ^{15}N-^{1}H sites, together with 62 ^{13}C and 103 ^{1}H backbone and sidechain resonances, were misassigned in this initial resonance assignment list. These misassignments did not show inconsistencies in the initial automatic backbone assignment process nor in the preliminary manual sidechain assignment efforts. At the time that the work was done (and prior to having the

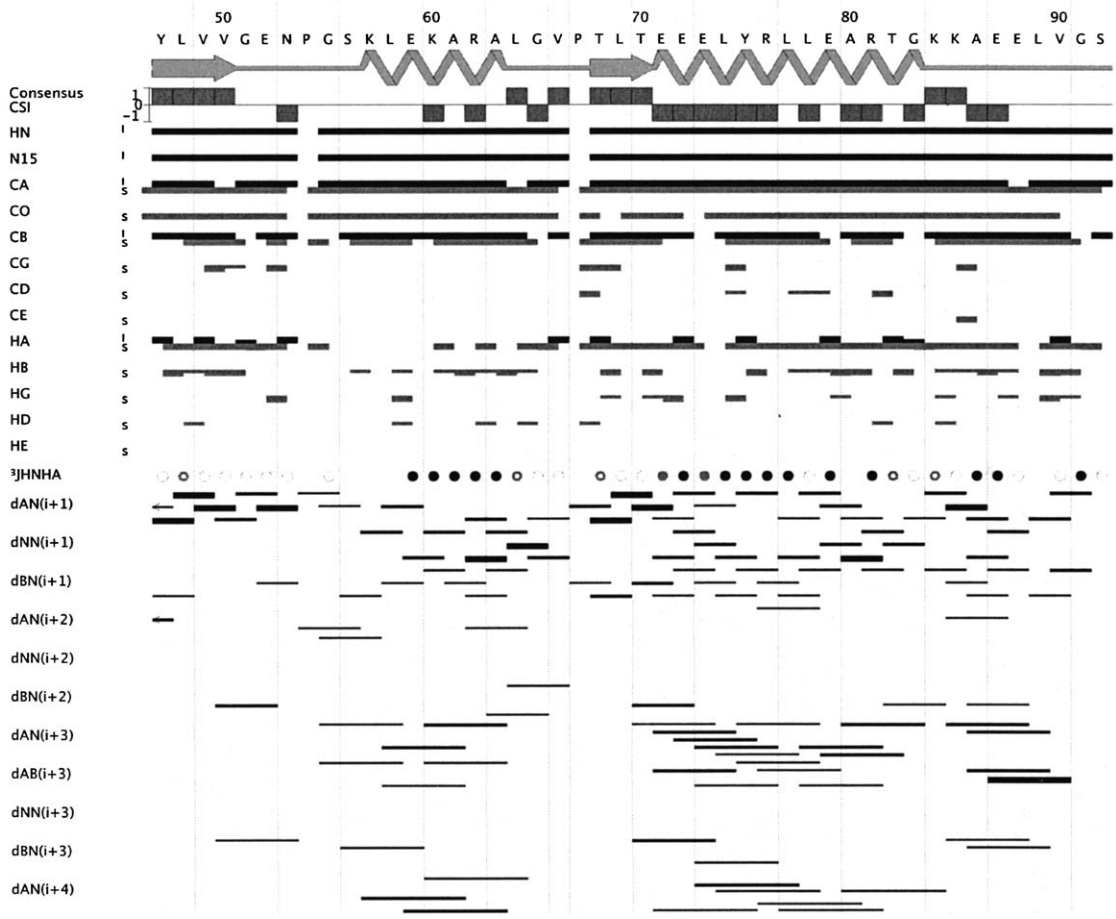

Figure 4. Portion of CMap Image of assignment completeness and consistency for the BRCT domain from *Thermus thermophilus* DNA ligase (Sahota et al., 2004; BMRB accession number 5328). The first row is the protein sequence. The next row annotates the secondary structure. The third row is the consensus chemical shift index (CSI) calculated from the Hα, Cα, Cβ and C′ chemical shifts, based on the method of Wishart et al. (1997). The next 13 rows summarize triple-resonance connectivity data. The next row summarizes $^{3}J(H^{N}\text{-}H^{\alpha})$ coupling data. The final 11 rows summarize sequential and medium-range NOE data validating the assignments and secondary structure.

AVS suite described here), several errors were fixed by careful manual sidechain assignment efforts. Still, some erroneous assignments were only detected during the analysis of the NOESY spectra and initial 3D structure calculations (Swapna et al., 2001).

Table 6 shows the summary of AVS analyses for the initial incorrect assignment list obtained for RbfA. In particular, both the **validate_assignment.pl** (V) and **typing_degeneracy.pl** (T) analyses directly detected 21 out of the 27 misassigned amide H^{N}-N chemical shift pairs. The 15 misassigned amides detected by **typing_degeneracy.pl** were labeled as 'possibly misassigned'. Of the 6 undetected amide misassignments, 4 are next to directly detected misassignments. The last 2 undetected amide H^{N}-N chemical shift pair misassignments are the correct assignments for 2 other detected misassignments. Also, these analyses directly detected 46 out of 62 ^{13}C chemical shift misassignments, and 60 out of 103 non-amide ^{1}H chemical shift assignments (including some aromatic resonance assignments). The lower detection rate for the ^{1}H chemical shifts comes from the larger expected ranges for these types of chemical shifts and the overlap in these expected ranges across the amino acid types.

Both analyses have low false positive rates (False T and False V in Table 6), in which spin systems flagged as 'suspicious' were indeed found to be correctly assigned. The **typing_degeneracy.pl** analysis had two false positives in the form of two different isolated spin systems with low linkage strength. All of the flagged spin system segments had some amide

Table 6. Summary of validation analyses of several protein assignments sets

Detection method	Amide H^N-N shifts	Carbon shifts	Hydrogen shifts
Intermediate RbfA			
T^a	15	30	43
V^b	10	28	34
T^a and V^b	4	12	17
T^a or V^b	21	46	60
Not detected[c]	6	16	43
Total present[d]	27	62	103
False T^a	2	–	–
False V^b	0	1	1
Final RbfA[e]			
T^a	0	0	0
V^b	0	0	0
T^a and V^b	0	0	0
T^a or V^b	0	0	0
Not detected[c]	0	3	1
Total present[d]	0	3	1
False T^a	0	–	–
False V^b	0	0	0
Final NESG ER14[f]			
T^a	0	0	0
V^b	0	0	0
T^a and V^b	0	0	0
T^a or V^b	0	0	0
Not detected[c]	0	0	0
Total present[d]	0	0	0
False T^a	6	–	–
False V^b	1	0	2
Final NESG JR19[g]			
T^a	0	0	0
V^b	0	0	0
T^a and V^b	0	0	0
T^a or V^b	0	0	0
Not detected[c]	0	0	0
Total present[d]	0	0	0
False T^a	5	–	–
False V^b	0	0	0

[a]Number of possibly erroneous assignments flagged by **typing_degeneracy.pl** Perl program analysis.
[b]Number of possibly erroneous assignments flagged by **validate_assignments.pl** Perl program analysis.
[c]Erroneous assignments not flagged by either **typing_degeneracy.pl** or **validate_assignments.pl** Perl programs.
[d]Total number of erroneous assignments in the chemical shift list.
[e]BMRB accession number for RbfA is 5093.
[f]BMRB accession number for ER14 is 5596.
[g]BMRB accession number for JR19 is 5691.

misassignments. The **validate_assignments.pl** analysis had only one false positive ^{13}C 'misassignment' and one false positive ^{1}H 'misassignment' which upon manual analysis was determined to be correct. This demonstrates that both analyses have significant discrimination power.

Table 6 also shows the summary of AVS analyses for three other protein assignment lists. These include a later stage AutoAssign assignment of RbfA, after significant cleanup and improvement of input peak lists, an AutoAssign assignment list for *E. coli* hypothetical protein yggU (NESG ER14), and an AutoAssign assignment of *P. horikoshii* ribosomal protein S28E (NESG JR19). In all three examples, the number of actual assignment errors is minimal. In fact, only the later stage automated RbfA analysis has three ^{13}C 'misassignment' and one ^{1}H 'misassignment' based upon a comparison with the published chemical shift list (Swapna et al., 2001) which had been refined by further manual analysis of the NMR spectra. The number of false positives for both **validate_assignments.pl** and **typing_degeneracy.pl** analyses are low. The false positive rate for **validate_assignments.pl** analyses is zero for all three of these examples using accurate assignment lists as input. The false positive rate for H^N-N amide analysis with **typing_degeneracy.pl** is zero, 6, and 5 for late stage RbfA, NESG ER14, and NESG JR19, respectively. It should be emphasized that the AVS analysis provides clues to the user of where assignments *may be incorrect.* In the end, it is up to the user to decide if the assignment made is adequately supported by the data. In these last three examples, all assignments, flagged as 'suspicious' or 'possibly misassigned' were determined to be correct upon careful manual inspection of the NMR data.

Discussion

The Assignment Validation Software suite has three major components (Table 2), the **validate_assignments.pl** Perl program, the **typing_degeneracy.pl** Perl program, and the AutoAssign **CMI Editor.** Other components of AVS support the use of these three major components (Table 1). Together, the AVS suite provides a set of tools that enables strict consistency checking of chemical shift assignments against the spectral data they are derived from and their expected values ('BMRB Expected Set').

A possible improvement in the AVS suite and other future validation algorithms is to directly handle the

deviations from normality of the 'BMRB Expected Set' due to secondary structure effects. This would entail an extensive analysis of correlations between secondary structure and chemical shift and require datasets with these correlations. Currently, we feel that the BMRB does not contain enough high quality, well-referenced datasets with the necessary correlations to allow such an analysis. Once enough data becomes available, such an addition would improve performance in the statistical evaluation and amino acid type analysis in the **validate_assignments.pl** Perl program. It would also improve the *segment mapping likelihood* and *segment similarity scores* used in the *segment mapping degeneracy analysis* in the **typing_degeneracy.pl** perl program.

The reports from the AVS suite are highly discriminating with low rates of false positives. In practice, these flagged assignments are simply those that require careful manual inspection. The AVS suite can handle data collected from fully and partial uniformly deuterium labeled samples. Also, the AVS suite can perform these analyses for assignments generated either with AutoAssign or with alternate automated or manual methods. The AVS suite can carry out these validations without the use of the protein structure, and can be used when NOESY and/or side chain resonance assignment data are unavailable, or while waiting on the completion of these experiments. The AVS suite analyses are also not sensitive to assignment completeness. Thus, the AVS suite provides tools for consistency checking throughout the assignment and structure determination process. Moreover, the **CMI Editor** allows visualization of all relevant data for final evaluation and publication of protein chemical shift assignments. The AVS suite is available at http://www-nmr.cabm.rutgers.edu.

Acknowledgements

We thank G.V.T. Swapna for helpful discussions and suggestions, and the intermediate RbfA assignments, James Aramini for ER14 and JR19 data and assignments, Daniel Monleon for WR33 assignments and spin system connectivity data, Marina Kiriyeva for Java coding related to the AutoAssign CMI Editor. We also thank Janet (Yuanpeng) Huang, David Snyder, and Michael Baran for useful comments on the manuscript. This work was supported by grants from the Protein Structure Initiative of the National Institute of Health (P50 GM62413) and The New Jersey Commission on Science and Technology (99-2042-007-13).

References

Brenner, S.E. (2001) *Nat. Rev. Gen.*, **2**, 801–809.

Constantine, K.L., Muller, L., Goldfarb, V., Wittekind, M., Metzler, W.J., Yanchunas, Jr., J., Robertson, J.G., Malley, M.F., Friedrichs, M.S. and Farmer, B.T. (1997) *J. Mol. Biol.*, **267**, 1223–1246.

Duda, R.O. and Hart, P.E. (1973) *Pattern Classification and Scene Analysis*, John Wiley and Sons, New York, NY.

Forsyth, D.A. (1984) *Isot. Organ. Chem.*, **6**, 1–66.

Gardner, K.H. and Kay, L.E. (1998) *Annu. Rev. Biophys. Biomol. Struct.*, **27**, 357–406.

Grzesiek, S. and Bax, A. (1993) *J. Biomol. NMR*, **3**, 185–204.

Güntert, P., Salzmann, M., Braun, D. and Wüthrich, K. (2000) *J. Biomol. NMR*, **18**, 129–137.

Hansen, P.E. (1988) *Prog. NMR Spectr.*, **20**, 207–255.

Herrmann, T., Güntert, P. and Wüthrich, K. (2002) *J. Mol. Biol.*, **319**, 209–227.

Huang, Y.J. (2001) *Automated Determination of Protein Structures from NMR Data by Iterative Analysis of Self-Consistent Contact Patterns*, Rutgers University, Dissertation.

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

Logan, T.M., Olejniczak, E.T., Xu, R.X. and Fesik, S.W. (1992) *FEBS Lett.*, **314**, 413–418.

Monleon, D., Chiang, Y., Aramini, J.M., Swapna, G.V.T., Macapagal., D., Gunsalus, K., Kim, S., Szyperski, T. and Montelione, G.T. (2003) *J. Biomol. NMR*, in press.

Montelione, G.T., Lyons, B.A., Emerson, S.D. and Tashiro, M. (1992) *J. Am. Chem. Soc.*, **114**, 10974–10975.

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

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

Moseley, H.N.B., Monleon, D. and Montelione, G.T. (2001) *Meth. Enzymol.*, **339**, 91–108.

Oldfield, E. (1995) *J. Biomol. NMR*, **5**, 217–225.

Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J. (1991) *J. Biomol. NMR*, **1**, 217–236.

Sahota, G., Sharon Goldsmith-Fischman, S., Dixon, B., Huang, Y.J., Aramini, J., Yin, C., Xiao, R., Bhattacharya, A., Monleon, D., Swapna, G.V.T., Anderson, S., Honig, B., Monteiro, A.N.A., Montelione, G.T. and Tejero, R. (2004) *Proteins: Structure Function Genetics* (submitted).

Swapna, G.V.T., Shukla, K., Huang, Y.J., Ke., H., Xia, B., Inouye, M. and Montelione, G.T. (2001) *J. Biomol. NMR*, **21**, 389–390.

Venters, R.A., Farmer II, B.T., Fierke, C.A. and Spicer, L.D. (1996) *J. Mol. Biol.*, **264**, 1101–1116.

Wishart, D.S. and Case, D.A. (2001) *Meth. Enzymol.*, **338**, 3–34.

Wishart, D.S. and Sykes, B.D. (1994) *J. Biomol. NMR*, **4**, 171–180.

Wishart, D.S., Watson, M.S., Boyko, R.F. and Sykes, B.D. (1997) *J. Biomol. NMR*, **10**, 329–336.

Xu, X. and Case, D.A. (2001) *J. Biomol. NMR*, **21**, 321–333.

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

Zweckstetter, M. and Bax, A. (2001) *J. Am. Chem. Soc.*, **123**, 9490–9491.