AutoStructure

RPF

Reference Manual
Version 2.1.1

Draft 3

We would like to thank Pavithra Sivakumar for helping in preparing this manuscript.

Y. J. Huang (yphuang@cabm.rutgers.edu)
G. T. Montelione (guy@cabm.rutgers.edu)
Center for Advanced Biotechnology and Medicine
Rutgers and Robert Wood Johnson Medical School

Contents

1. Introduction

1.1 AutoStructure – A topology constrained distance network algorithm for protein structure determination from NOESY data

AutoStructure is a protein structure determination tool that uses uninterpreted NOESY cross peaks together with structure calculation programs like XPLOR or DYANA to generate a 3D structure of the protein that is as close to the true structure as possible. AutoStructure uses an iterative bottom-up topology-constrained approach to analyze NOE peak lists. It first builds an initial fold based on intraresidue and sequential NOESY data, together with characteristic NOE patterns of secondary structures, including helical medium-range NOE interactions and interstrand β-sheet NOE interactions, and unique long-range packing NOE interactions based on chemical shift matching and symmetry considerations. Unassigned NOESY cross peaks are not used in structure calculations. Additional NOESY cross peaks are iteratively assigned using intermediate structures and the knowledge of high-order topology constraints of α-helix and β-sheet packing geometries. This protocol, in principal, resembles the methodology that an expert would utilize in manually solving a protein structure by NMR.

For citing AutoStructure you may consider:

Huang, Y. J.; Tejero, R.; Powers, R.; Montelione, G.T. A topology-constrained distance network algorithm for protein structure determination from NOESY data. *PROTEINS: Struct. Funct. Bioinformatics* 15, 587-603 (2006)

1.2 RPF – Protein NMR structure quality assessment tool

RPF uses a novel, rapid, and simple approach for calculating global NMR structure quality scores. This program calculates RECALL, PRECISION, and F-MEASURE (RPF) scores assessing how well the query 3D structure(s) fit to the experimental NOESY peak list and resonance assignment data. RPF scores quickly assess the goodness-of-fit of the query structure(s) to these experimental data, and can be used as a guide for further structure refinements.

RPF also calculates discrimination power (DP) scores, which estimate the difference in F-MEASURE scores between the query structure and “random coil” structures, as an indictor of the correctness of the overall fold. The program is

useful for quality of control protein NMR structures determined by automated or manual methods.

For citing RPF you may consider:

Huang, Y. J.; Powers, R. & Montelione, G. T. Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. *J Am Chem Soc* 127, 1665-74 (2005)

2. Installation

The following steps are required:

1. Install programs "rasmol, dyana or xplor, DQS or PBS" and perl-tk modules on your computer/clusters.

```
DQS and PBS are batch queuing systems used to manage
clusters of PCs. If you are using clusters, please contact
your system admin for installation.
```

2. Update the environment variables defined the script files `bin/autostructure', `bin/asgui' and `bin/CreateProc'

`bin/autostructure' - run AutoStructure without GUI
`bin/asgui' - run AutoStructure with GUI
`bin/CreateProc' - run Structure Calculations using Xplor or Dyana

3. Update the following environment variables defined in the scripts `bin/autostructure', `bin/asgui' and `bin/CreateProc':

For autostructure and asgui

RASMOL	rasmol command (viewer command, optional)
ASPATH	path of AutoStructure directory
ASBIN	path of AutoStructure bin directory or use default
NALIB	path of NOESY_Assign library (a submodule of AutoStructure) or use default
AyudaPATH	path of PDBSTAT library (a submodule of AutoStructure) or use default
PERL5LIB	path of the perl-tk GUI (used by `asgui' only) or use default

For CreateProc

XplorCommand	xplor command for structure calculations
DyanaCmd	dyana command for structure calculations
SubmitDQS	command to run structure calculations if using DQS system
SubmitPBS	command to run structure calculations if using PBS system
PDBStat	command to run PDBStat (a submodule of AutoStructure)

3. Using AutoStructure

3.1 Modules

AutoStructure has the following three sub-modules:

NOESY_Assign	make NOE assignments and generate distance and h-bond constraints
HYPER	generate dihedral angle constraints
CreateProc	submit signal or parallel DYANA/XPLOR structure calculations
PDBStat	tools to analyze 3D structures

3.1.1 Methods used by NOESY_Assign

3.1.1.1 Initial fold methods (cycle 1-0)

Rule-in methods for NOE assignments are:

- **EXPECT**: consistent with intra-residue, sequential, and secondary structure contact patterns
- **SYM**: one-to-one mapped unique symmetric assignments with high potential contact confidents.
- **UNIQUE**: unique and with high potential contact confidents.
- **INITF**: assigned based on the initial input structures

Rule-out methods

- **DEL_SS** - inconsistent with secondary structures
- **DEL_LOC** (Chemical shift refinement): for atoms that have small chemical shift variations (observed from NOE assignments made by EXPECT method), a tighter match tolerance is used. Matches outside this tighter match tolerance are ruled out.

3.1.1.2 Iterative fold methods (other cycles)

Rule-in methods (used in other cycle *-0):

- **PDB**: consistent with 3D structures
- **SYM**: symmetric NOE assignments that best supported by the model structures.

- **BEST**: NOE assignments that best supported by the model structures.
- **DEL_PDB/DEL_TOP/DEL_ISO**, etc: unassigned by other reasoning based on the model structures

Rule-out methods (used in validation cycle)
- **VIO** – distance constraint violated with 3D structures

3.2 Using the Interface

3.2.1 Prepare the input data:

AutoStructure requires the following input for structure determination –
a. Amino acid sequence of the protein
b. List of resonance assignments
c. List of multidimensional NOESY cross peak frequencies and intensities
d. Optional list of scalar coupling data
e. Optional list of slow amide exchange data
f. Other available constraints such as residual dipolar coupling, disulphide-bond, distance and dihedral-angle constraint data,

and

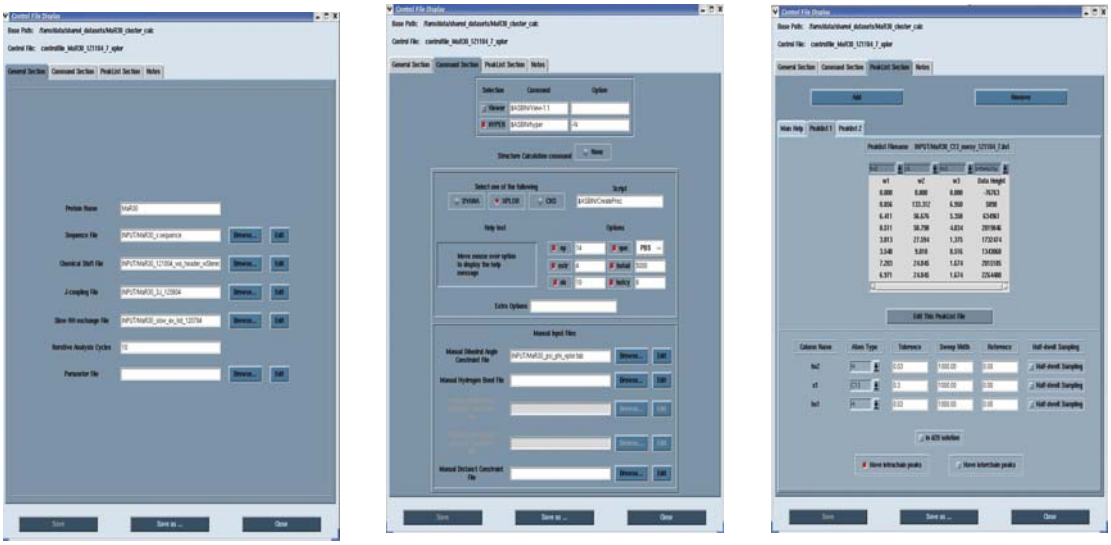
e. A control-file which instruct AutoStructure to read in all the input data and specify how to run structure calculations. The control-file can be generated using the GUI interface ("**File->New Control File**").

The general section of the GUI interface (Fig. 3.1) lets the user specify the input for an AutoStructure run. The command section lets the user select structure calculation commands (XPLOR or DYANA), and also optional input files such as manual dihedral angle constraint files or hydrogen bond files. The peaklist section allows the user to create or modify existing peaklists.

The formats for these input data and control-file are summarized in Appendix I.

Figure 3.1 - the control-file GUI interface

(a) general section (b) command section (c) peaklist section



3.2.2. Start AutoStructure Calculations

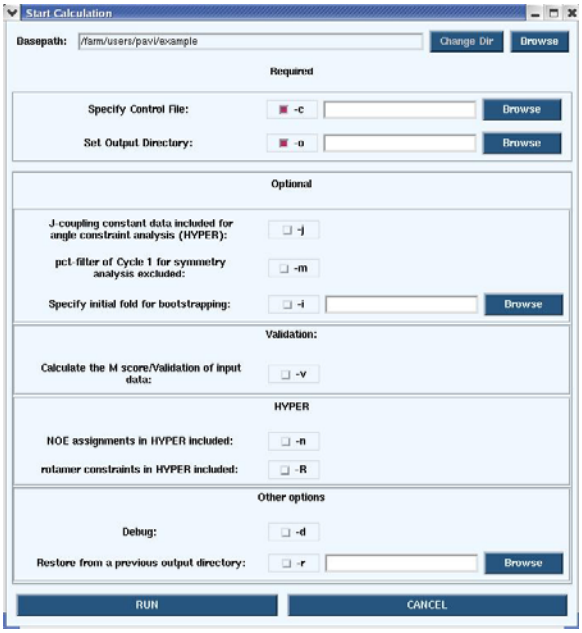


Figure 3.2

After specifying the location of the control file, the user can run AutoStructure (**"AutoStruc-ture -> Calc -> start"**) (Figure 3.2). The results are stored in the output directory the user specified, and can be analyzed using Auto-Structure's interface. You can also run the program in outside of the GUI interface. Type "**bin /auto-structure**" for instructions.

3.2.3. AutoStructure results

The results can be view graphically ("**AutoStructure->Open AS OutputDir**"). The taxt outputs are summarized in Appendix II.

a. Summary of results

The summary of results provides a plot view of the number of NOE assignments and the percentage of peaks assigned during each cycle in AutoStructure (Figure 3.3). The information is taken from the .ovw file generated by AutoStructure. As seen in the Figure 3.3, the percentage of peaks assigned in cycle 1 (60%) increases to 90% by the final cycle. A detailed view of the results from each cycle allows the user to see the number of backbone, side chain, intra-residue, sequential, medium-range, and long-range peaks that were assigned. The user can also compare the results to that of the previous cycle to see the new peaks that were added and the ones that were removed.

(a) (b)

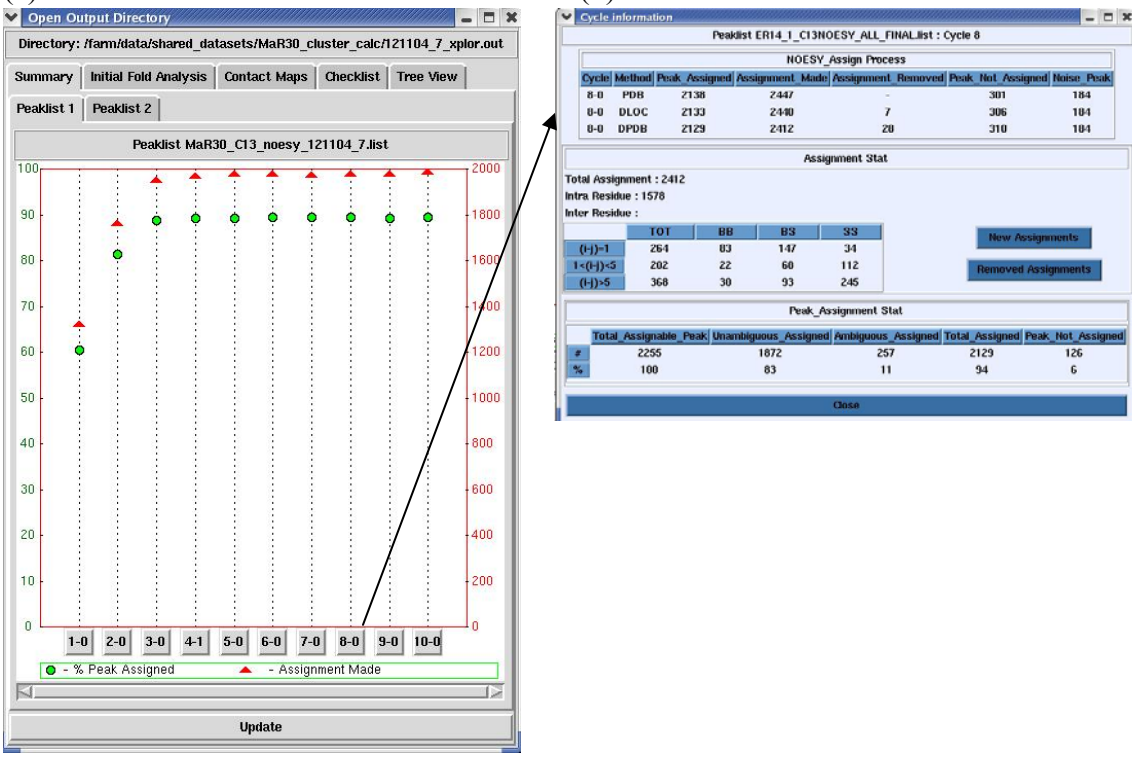


Figure 3.3 *– (a) The interface presents a plot view of the summary of peak assignments made during each cycle of the structure determination process in AutoStructure. The green dots represent the cumulative percentage of peaks assigned in each cycle, and the red traingles*

represent the cumulative number of NOE assignments made in each cycle. (b) The interface also gives a detailed description of the assignments made in each cycle, including fraction of inter-residue, sequential, backbone and side chain assignments made.

b. Initial Fold Analysis

The secondary structure analysis plays an important role in AutoStructure. The initial fold analysis view (Figure 3.4) presents the secondary structure analysis results from AutoStructure during its first cycle (based on the *sec report). Secondary structure mappings as dictated by CSI (Chemical Shift Indices), combined analysis of CSI, scalar coupling and the NOE data, and the final secondary structure results determined by AutoStructure are presented. Data from slow amide exchange and scalar coupling is also incorporated in this view. The user is also allowed to import the secondary structure prediction from the PROF server (http://www.aber.ac.uk/~phiwww/prof/) for comparison. An interface to submit PROF job is also implemented ("**MiscTools->Secondary Structure Prediction**").

(a) (b)

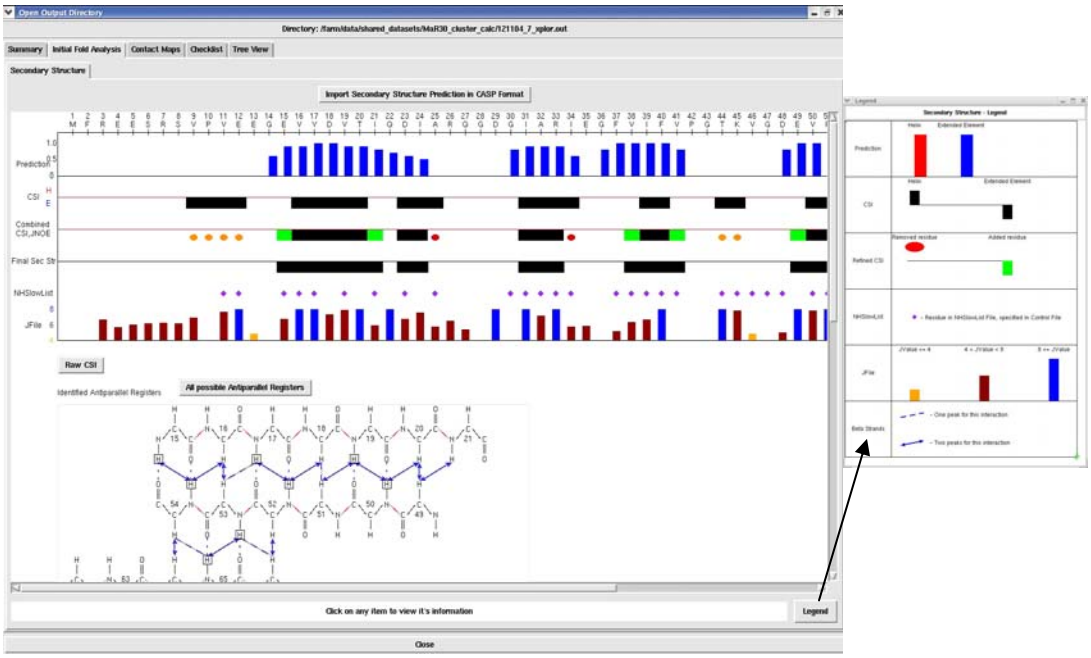


Figure 3.4 *– (a) The Initial Fold Analysis view provides a summary of the secondary structure analysis done by AutoStructure during the first initial fold analysis cycle. Predictions of secondary structure elements from chemical shift indices, scalar coupling data, slow amide exchange data, and combined secondary structure analysis resulting from chemical shift index,*

scalar coupling and NOE data are shown. The user can also import secondary structure predictions from the PROF server. (b) The legend for the secondary structure analysis results.

The ß-sheet topology of the protein in question is displayed along with the registers that were identified (Figure 3.5). All the antiparallel and parallel strands are displayed as identified by the algorithm, along with the registers and interactions that led to the identification of the ß-sheet. NOE interactions that are expected to be seen in the ß-sheet topology are shown in grey in the background, while the identified NOE interactions are superimposed on them in blue. Information on the peaks that led to the interactions is also available. This graphical display of the interactions in the ß-sheet allows the user to review the peak assignments associated with them, and to manually identify and refine their peak lists. You can also query all possible registers from the interface.

Validating the init fold analysis report from the GUI interface or with *sec file is recommended.

(a) (b)

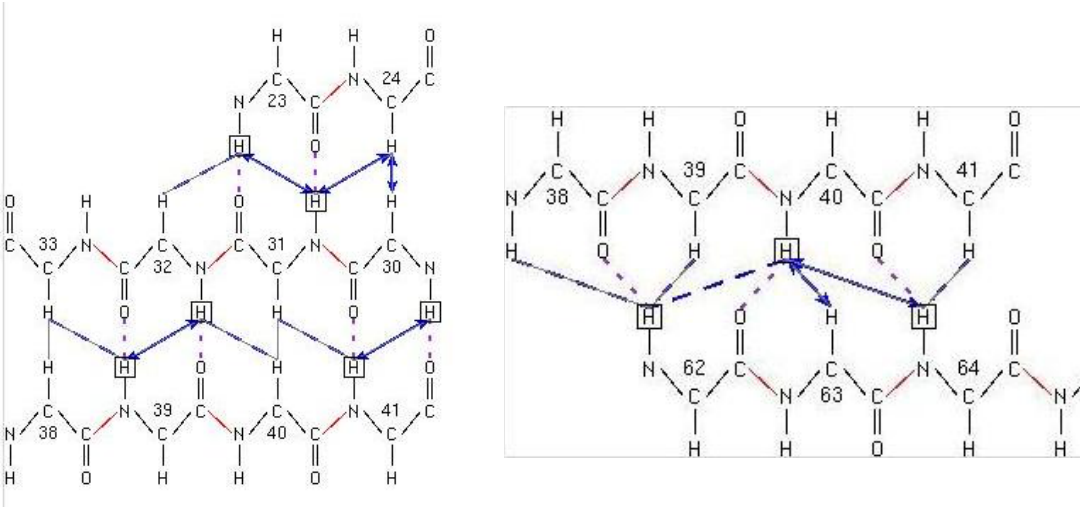


Figure 3.5 *– The initial fold analysis view of the interface also shows the ß-sheet registers identified by AutoStructure. The expected interactions are shown in grey, while the observed interactions are shown as blue arrows. (a)The antiparallel registers identified are shown (b) The parallel registers that were identified are shown. The peak information for each of the observed interactions is also available.*

c. Contact Map display

The contact map display (Figure 3.6) in the AutoStructure interface presents a grid that maps all the interactions between any pair of residues. Backbone interactions are shown in red in the lower triangular region of the diagonal. All interactions (backbone and sidechain) are shown in the upper triangular region in black. Intraresidue interactions are along the diagonal in green. The peaks and distances corresponding to interactions can also be viewed in tabular form by querying for interactions between any two variable residues.

(a) (b)

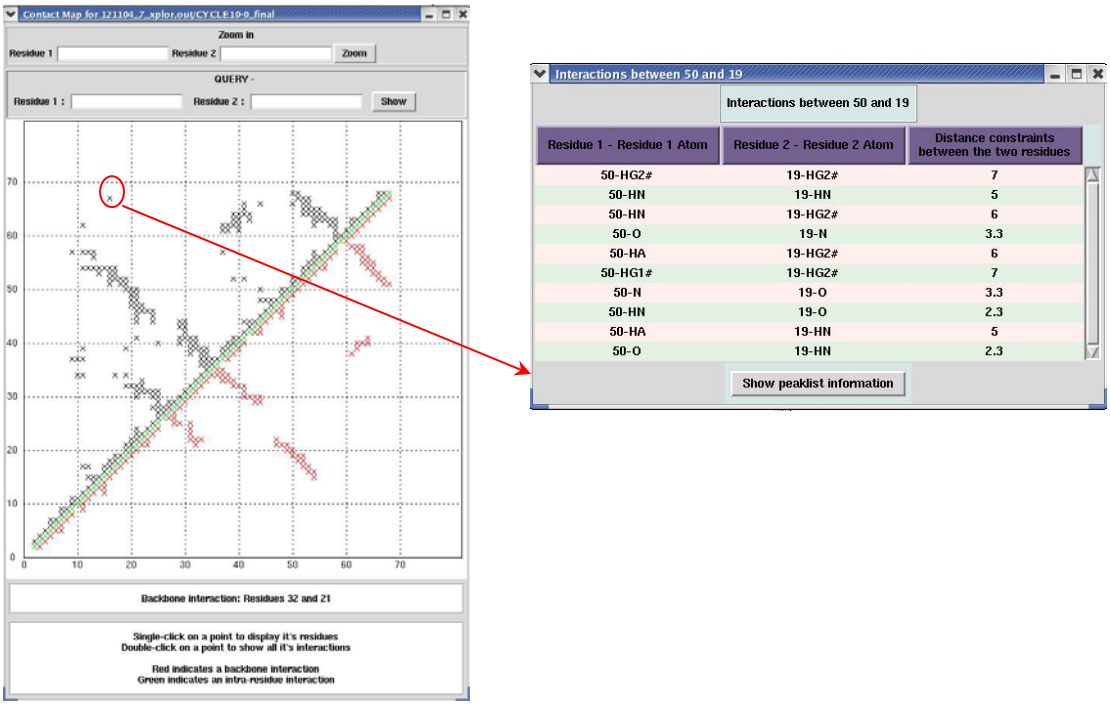


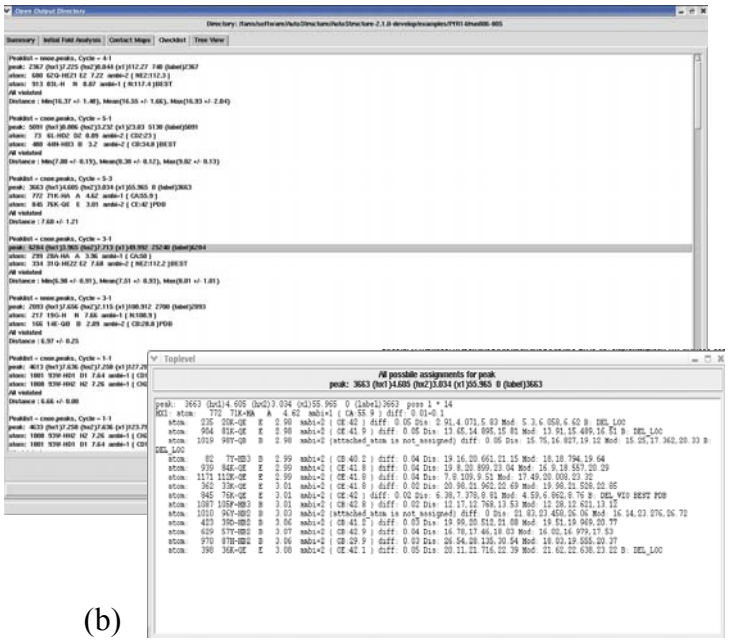
Figure 3.6 *– (a) The contact map display shows a grid of interactions between pairs of residues in the protein. (b)A list of all the interactions between any two residues is also shown in tabular view.*

d. Checklist

This section of the initial fold analysis summary reports a list of those peaks that cannot be explained by the final structure but were used in the intermediate analysis stages (Figure 3.7). The "unassigned" peaks are a subset of the false negative interactions (interactions that are part of the input peaklist, but cannot be explained in the final structure) that were used by the algorithm in the intermediate stages of structure determination. Each peak is displayed along with its peaklist, cycle, and distance information, along with the reason it was assigned

and then unassigned. All possible assignments of each peak can also be viewed in tabular form. **This list of unassigned peak could include potential noise peaks or incorrect assignments, and is thus useful in refining the input NOESY peak lists.**

(a)



(b)

Figure 3.7 *– (a) The checklist section presents a list of peaks that were unassigned in the final structure but were used in the intermediate analysis. (b)A tabular view of all possible assignments of the peaks is also shown.*

3.3 Other applications

AutoStructure can be used at various stages of resonance assignments for validation. For example, given the backbone resonance assignments and 3D 15N-NOESY peak lists, AutoStructure can assign all backbone related intra and sequential NOEs and identify all secondary structure elements. These backbone related intra and sequential NOE connectivities are commonly used for cross-validation of backbone sequential connectivity derived from triple resonance methods. Given near complete backbone and side-chain resonance assignments and 3D HCCH-COSY peak lists, AutoStructure can assign all peaks in the 3D HCCH-COSY peaks for validation of the two-bond and three-bond connectivity of the side-chain resonances.

3.4 Frequently Asked Questions

3.4.1. Requirements for resonance assignment table

AutoStructure uses a chemical shift index method for secondary structure analysis and therefore requires accurate chemical shift referencing for Cα, and Cβ resonances. This chemical shift index method relies on the use of the recommended IUPAC chemical shift referencing method with DSS as the reference compound. High quality AutoStructure calculations require the input resonance assignment table to be more than 85% complete. For each aromatic residue, at least one aromatic side chain proton should be assigned in order for AutoStructure to define its ring packing.

3.4.2. Requirements for NOE peak lists

Peak lists do not have to be perfect. AutoStructure can handle the presence of artifactual peaks and incompleteness; however, inaccurate or imprecise peak picking can considerably limit the performance of the program. Intense solvent lines, ridges and/or sinc wiggles should be manually inspected and remove from the peak lists. Many NOE peaks may overlap with solvent lines and become hard to peak picking. However, collecting 3D 13C-NOESY in D20 can minimize such problem. AutoStructure can handle aliased/folded peaks. High quality AutoStructure calculations require the input peak list to contain at least 90% real cross peaks.

3.4.3. Requirements for matching the NOE peak lists and resonance assignments

AutoStructure calculates an M-score which estimates the percent of predicted conformation-independent two- and three-bond connected NOE-linked proton pairs that are missing from the NOE peak lists. **Select "–v" when start AutoStructure calculations.**

Four factors can contribute to high M scores: i) misalignment between chemical shifts from NOE peak lists and the resonance assignment table; ii) significant differences in the digital resolutions between chemical shifts from NOE peak lists and the resonance assignment table; iii) poor quality of NOE peak lists; iv) incorrect resonance assignments. A high M score (i.e. > 25%) suggests that at least one of the input data sets (R and/or NOE) are of inadequate quality and need to be improved. Those predicted two- and three-bond connected NOE-linked

proton pairs missing from the NOE peak lists are reported (in file *_NA.QM) for the user to improve the corresponding chemical shift assignments, and/or identifying the expected NOESY cross peaks in the corresponding NOESY spectrum.

*QM	Report M scores, calculated using -v. It provides guidance for chemical shift refinement. For atoms that shifts consistently more than 0.3ppm for C/N or 0.03ppm for H, it is recommended to adjust them manually.

AutoStructure requires that all NOESY spectra be accurately referenced relative to the values of chemical shifts reported in the resonance assignment table. For each frequency dimension, the software computes the overall average chemical shift match difference from these predicted NOE-linked proton pairs (in the *_NA.ovw file). Consistent spectral referencing is achieved using these differences as global reference correction factors for the target spectrum, providing a tighter match between NOE peak lists and resonance assignment table, and allowing the use of smaller matching tolerances Δ for further NOESY interpretation.

Example output from ER14_NA.ovw:

```
# Summary for ER14_1_500_3DN15NOESY_FINAL.list
Total simulated peaks: 125 Number of peaks NOT matched in the Peak List: 4
M score = 0.032

Average shift in HX: -0.0056 Average ab-shift in HX: 0.0062(Assignment used: 110)
Average shift in X: 0.007 Average ab-shift in X: 0.066(Assignment used: 100)
Average shift in HX2: -0.016 Average ab-shift in HX2: 0.016(Assignment used: 115)

# Summary for ER14_1_C13NOESY_ALL_FINAL.list
Total simulated peaks: 809 Number of peaks NOT matched in the Peak List: 52
M score = 0.064

Average shift in HX: -0.0089 Average ab-shift in HX: 0.0098(Assignment used: 346)
Average shift in X: -0.051 Average ab-shift in X: 0.065(Assignment used: 272)
Average shift in HX2: 0.0027 Average ab-shift in HX2: 0.0075(Assignment used: 399)
Overall M-score: 0.0599572 56/934
```

Average shift is the average of the summation of all matching differences. Average ab-shift is the average of the summation of the absolution values of matching differences.

3.4.4 Refining chemical shifts based on AutoStructure reports

Step 1. Do one cycle of AutoStructure calculation with option **–v**

Step 2. Check file *_NA.ovw. It report the M-score and average chemical shift variations for each frequency dimension. Minimize the average chemical shifts variations for each peak list and rerun the calculation (GUI interface will be developed to view this section)

Step 3: AutoStructure updates chemical shift assignments within each assignment cycle based on the updated NOE assignments. You may use **RefinedShift.bmrb** from CYCLE* output directory as a guide for refining chemical shifts.

You may also use other output files from AutoStructure:

Check files *.QM and *.exm (GUI interface will be developed to view these two files)

File *.QM have two parts:
1). Part1: give a list of expected two- or three-bond connected NOEs that are missing from the peak list. Check these Part1 results with the spectrum. If missing from the spectrum -> there may be an incorrect resonance assignment
2). Part2: report the chemical shift variations for every atom with its matched expected two- or three-bond connected NOEs. Atoms that are consistently off by > 0.03ppm H or 0.3ppm C/N in more than two matched expected NOEs are recommended for manual refinements

Two parts of file *.exm are useful:
1). Part1: Unique method summary
2). Part3: EXPECT method summary which give a list of matched expected two-, three-, four-bond connected NOE assignments that did not assigned by EXPECT method (off by > 0.03ppm in H or 0.3ppm in C/N). Atoms that are consistently off by > 0.03ppm H or 0.3ppm C/N in more than two matched expected NOEs are recommended for manual refinements.

3.4.5 Tolerances

Match tolerances: H 0.05, C/N 0.5ppm

Assignment tolerances for initial fold analysis (EXPECT, SYM, UNIQUE method): H 0.03, C/N 0.3ppm
Assignment tolerance for iterative fold analysis (PDB, SYM, INITF):

- proton pairs with <5 Å: H 0.05, C/N 0.5ppm
- proton pairs with >5 Å : H 0.03, C/N 0.3ppm

How to change tolerances for a specific data set?

- Match tolerances -> control-file
- Assignment tolerances -> par.tbl

3.4.6 Cross-strand alignment analysis

The registers are identified based on their interstrand frequency (total number of expected interstrand NOEs) in the peak list. The cutoff is 6 interstrand backbone NOEs for both antiparallel and parallel sheet alignment.

When there are noises in the peak list, noise peak can increase the frequencies of incorrect registers. Verifying all alignments' assigned long-range NOE is recommended. We have noticed that the noise in HN-HN region disturbs the analysis. A clean HN-HN region helps a lot. Other regions such as HA-HA, and HA-HN are also important.

If a warning is given as the following example:

```
*** WARNING (CASE(a)): Conflict for strands 73-81 and 73-116
```

this warning indicates that there are two possible alignments are inconsistent and have same scores, the program can not distinguish them. Solution can be: increase frequency cutoff in par.tbl, and or remove noise peaks

3.4.7 Disulfide bond and Cis-proline

Potential cis-peptide bonds and disulfide bonds are identified and reported to the user for expert analysis and validation. You can also specify them in the control-file.

3.5. Common Errors

3.5.1 Can not open input file */Workingcycle/*.pdb

WorkingCycle is the intermediate directory used for XPLOR/DYANA calculations. When there is an error during the XPLOR/DYANA calculation, it is reported in the sa_for_AS_*.inp.out (xplor) or log* (dyana) files

When No *.pdb files are generated and the program will report

```
Can not open input file */WorkingCycle/*.pdb
```

Possible reasons:

- Some atoms names are not BMRB standard
- Some residues defined in the sequence is different from the residue specified in the resonance assignment table
- Errors in the manual input constraint file

Solution:
- Check the log file in */WorkingCycle/
- Correct the input files accordingly

When AutoStructure stops normally, it will print 'The program is finished.' at the end of the *.ovw file. Sometimes, when the queue system is unstable, the calculation may be stopped by the queue system or hung. In these cases, 'The program is finished.' is not printed in the *.ovw file.

4. Using RPF

4.1. RPF and DP scores

4.1.1 Definition

Table 4.1

NMR Data	Peak observed	Peak not observed
Interaction retrieved by Query Structures	TP	FP
Interaction not retrieved by Query Structures	FN	TN

There are four possible outcomes from the comparison of the query structures to the original peaklist:

- True Positive interactions are those observed both in the peak lists and final 3D structures
- True Negative interactions are those that are neither observed in the peak lists nor in the 3D structures
- False Positive interactions are those that are present in the 3D query structure but not present in the peak lists
- False Negative interactions are those peaks observed from the experimental data set that are not part of the 3D structure.

$$\text{Recall} = \frac{TP}{TP + FN} \qquad \text{Precision} = \frac{TP}{TP + FP} \qquad \text{F} - \text{measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

- Recall measures the percentage of peaks that are retrieved by the algorithm and are thus part of the query structure.
- Precision measures the fraction of retrieved proton pair interactions in the query structure whose back-calculated NOE peaks are part of the original peak list.
- F-measure which takes both Recall and Precision into account reflects the overall performance score of the structure.

A Discriminating Power (DP) score, which is a normalized F-measure statistic, is also developed to account for lower-bound and upper-bound values of the F-measure that are indicated by the NMR data quality. $0 \leq DP \leq 1$.

4.1.2 Applications

Comparing Recall and Precision scores during the course of a structure refinement can help to improve the peak picking process and/or identify errors in the input data, allowing refinement of the input used in the structure determination process. Generally, a reduced Recall rate compared with the Precision rate may suggest the existence of "noise peaks" in the input data set. High Recall rate compared with the Precision rate suggests that some weak NOE cross peaks have not been included in the NOESY peak lists because the corresponding signal-to-noise ratios are low. Good quality structures should have high Precision rates (few short inter-proton distances that do not have corresponding NOEs in the peak lists). Factors that could cause low Precision scores include surface amide proton saturation transfer, solvent exchange broadening, and conformational exchange broadening.

The F-measure score provides a good measure of the overall fit between the query structure and the experimental data, while the DP score measure how the query structure is distinguished from a freely rotating chain model, accounting for data quality. Low F-scores indicate that the structure does not fit well with the input data. High F-scores and low DP-scores indicate that the NMR data does not have enough long-range information that can distinguish the structure from a freely-rotating chain model. **Structures with F-measure > 0.9 and the DP score > 0.7 correlates to structures having accuracies of < ~ 2 Å rmsd.**

RPF scores are rapid and easy to compute, and are therefore well suited for routine use in quality assessment of NMR structures at different stages of analysis, using either manual or automated structure determination methods. At the initial stages of structure determination, they can assess the correctness of the protein fold and also guide the use of intermediate structures in making additional NOESY peak assignments. The false positive distribution analysis is useful in identifying inaccurate regions of the query protein structure, and can be used for further refinement and comparison to alternative structures generated from the same NMR data.

4.2 Using the Interface

The RPF interface provides a useful interface for the user to calculate RPF scores. Structures determined by manual or automated analysis, homology modelling, or X-ray crystallography can be used for RPF scores calculations.

4.2.1 Prepare the input

The following files are needed:

- NOESY peak lists
- Resonance assignment table
- Sequence file
- Control-file
- 3D structure coordinates in one pdb file

How to prepare input files are discussed in Appendix I. Control-file used for AutoStructure calculation can be used here for RPF calculations. The GUI interface for generating control-file is described in Section 3.2.1.

RPF is also specialized to calculate a set of RPF scores from the output of AutoStructure (Section 4.2.2).

4.2.2 Start the calculation

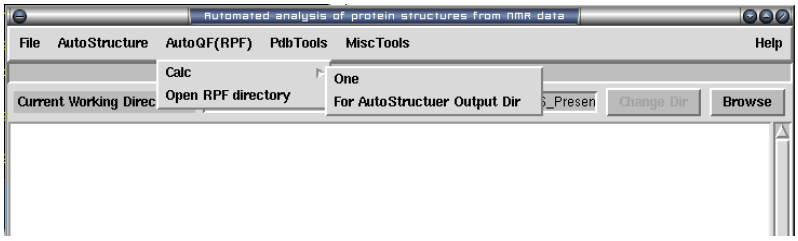


Figure 4.1

for one pdb file – select from menu: **AutoQF(RPF)->Calc->One**
for AutoStructure Output directory – select from menu: **AutoQF(RPF)->Calc->For AutoStructure Output Dir**

4.2.3 Open the output

Select from menu: **AutoQF(RPF)->Open RPF directory**

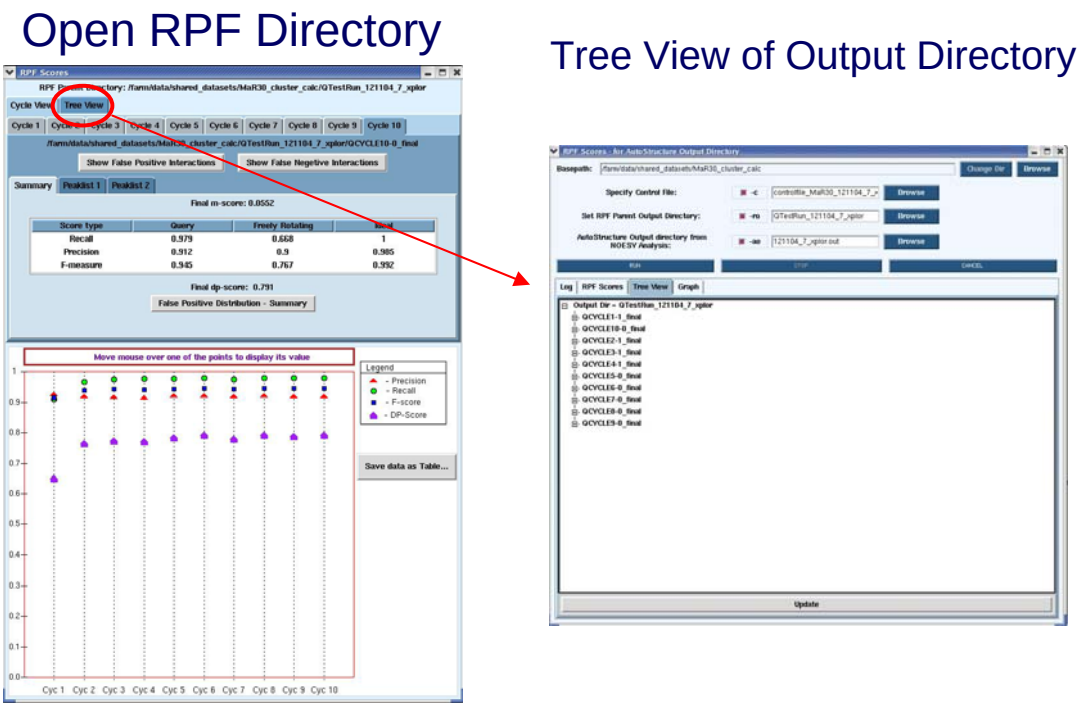


Figure 4.2

4.2.4 Quality control for iterative cycle analysis using the output of AutoStructure

RPF scores can be used as a quality control in iterative cycle analysis of protein structure determination in NMR. The RPF interface (Figure 4.2) displays the results of the iterative cycle analysis of NMR structure determination as a plot of the RPF and DP scores. The significant increase in the DP score from cycle 1 to cycle 10 demonstrates the improved accuracy of the final structure when compared to the initial and intermediate cycles. **By the final cycle the F-measure is > 0.9 and the DP score is > 0.7 which correlates to structures having accuracies of < ~ 2 Å rmsd.** During the iterative refinement process, as long as the structure does not have many bad proton-proton packing interactions, the Precision rate should be high and stay relatively constant. Figure 4.2 shows that Precision rates decrease slightly during the iterative process. This is due to the increased compactness of the structure over the course of the refinement, when additional weak NOE cross peaks predicted are missing from the input NOE peak lists (False Positives). The small decrease in precision over the course of refinement is diagnostic of the quality and completeness of the input NMR data.

4.2.5 False Positive Distribution

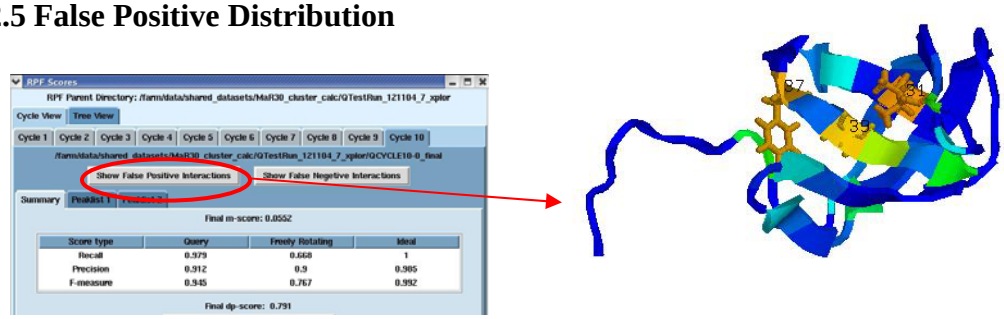


Figure 4.3 *- shows the False Positive distribution in a protein as presented by the RPF interface. The color coded regions in the 3D structure represent areas with False Positive interactions (interactions in the query structure with no corresponding peaks in the input peak lists).*

RPF maps the distribution of false positive interactions into the query structures. False Positive interactions are those interactions that are present in the final query structure but not part of the input NOE peak lists.

Precision rates measure the fraction of NOE interactions predicted by the structures and are also observed in the input NMR data. Thus, a higher the Precision corresponds to a lower number of false positive structural features. The graphical tool RASMOL is used to display the ribbon diagram of the query protein structure with color coded showing the missing interactions ranging from red (most problematic) to blue (least problematic) (Figure 4.3). The interface also provides a tabular view of the detail interactions given two residue numbers.

A Sparky peak list can be generated from these false positive interactions. Chemical shifts are generated from the resonance assignments. These false positive interactions can be queried with a query tool as shown in Figure 4.4.

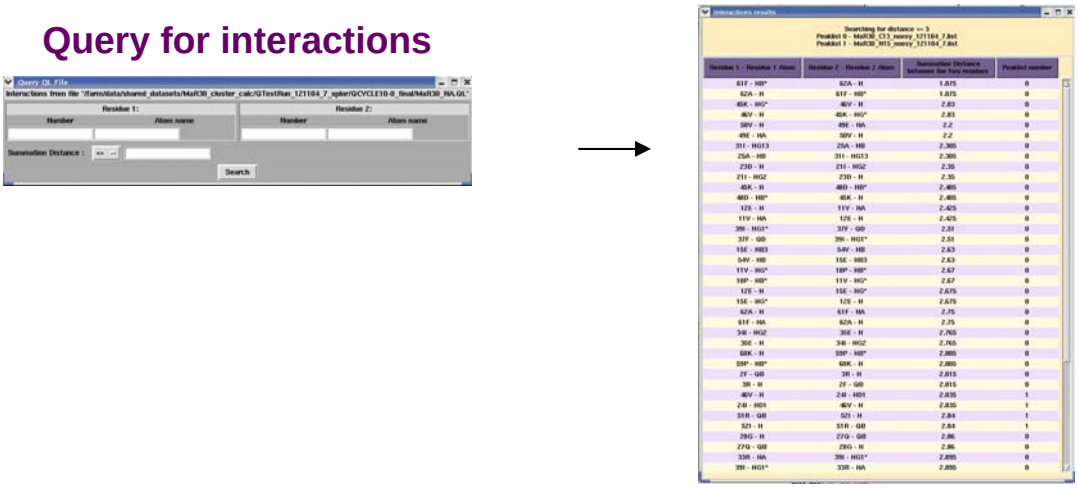


Figure 4.4 *The query interface for False Positive interactions*

4.2.6 False Negative Interactions

The interface for RPF also provides a display for those interactions that are present in the original input NOE peak list that do not have corresponding interactions in the final 3D structures. These interactions can then be used to evaluate the quality of the structure or the input NOE data. A Sparky peak list can be generated from these false negative interactions.

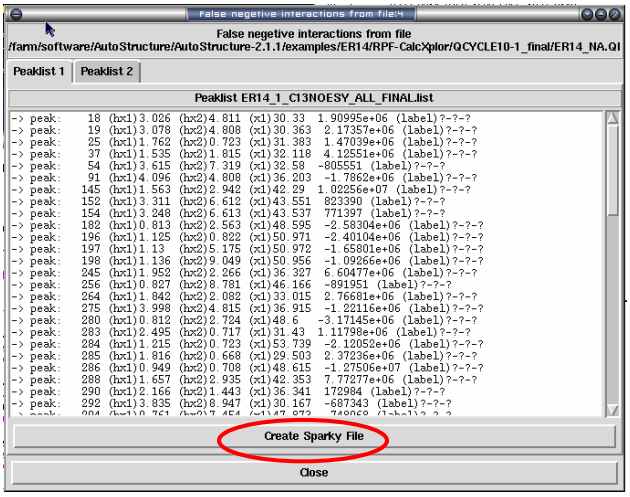


Figure 4.5 The interface showing all false negative interactions.

Appendix I Input file formats for AutoStructure/RPF

I.1 Sequence file – in BMRB NMR-STAR 2.1 format:

```
# TMZIP
#   _Residue_seq_code
#   _Residue_author_seq_code
#   _Residue_label

1  @  GLY  2  @  ALA  3  @  GLY  4  @  SER  5  @  SER
6  @  SER  7  @  LEU  8  @  GLU  9  @  ALA  10  @  VAL
11  @  ARG  12  @  ARG  13  @  LYS  14  @  ILE  15  @  ARG
16  @  SER  17  @  LEU  18  @  GLN  19  @  GLU  20  @  GLN
21  @  ASN  22  @  TYR  23  @  HIS  24  @  LEU  25  @  GLU
26  @  ASN  27  @  GLU  28  @  VAL  29  @  ALA  30  @  ARG
31  @  LEU  32  @  LYS  33  @  LYS  34  @  LEU  35  @  VAL
36  @  GLY  37  @  GLU  38  @  ARG
```

I.2 Resonance assignment table – in BMRB NMR-STAR 2.1 format

```
#
#Atom   Residue
#shift  Seq     Residue Atom    Atom    Shift/  Error/  Ambiguity
#assign code    Label   Name    Type    ppm     ppm     Code
#-----------------------------------------------------------------
#
2       1       GLY     CA      C       43.1    .       1
3       1       GLY     HA2     H       3.91    .       2
4       1       GLY     HA3     H       3.91    .       2
5       2       ALA     CA      C       52.5    .       1
6       2       ALA     CB      C       18.9    .       1
7       2       ALA     H       H       8.81    .       1
8       2       ALA     HA      H       4.41    .       1
9       2       ALA     HB      H       1.46    .       1
10      2       ALA     N       N       124     .       1
11      3       GLY     CA      C       45.1    .       1
12      3       GLY     H       H       8.73    .       1
13      3       GLY     HA2     H       4.05    .       2
14      3       GLY     HA3     H       4.05    .       2
15      3       GLY     N       N       108.9   .       1
16      4       SER     CA      C       58.4    .       1
17      4       SER     CB      C       63.4    .       1
18      4       SER     H       H       8.41    .       1
19      4       SER     HA      H       4.51    .       1
20      4       SER     HB2     H       3.96    .       2
21      4       SER     HB3     H       3.89    .       2
```

I.3 Peak lists – in table format (like sparky format)

```
#ID     Label   HX1     HX2     X1      Intensity
4       7       9.11    1.99    129.71  +6.82e+04
2       7       9.12    5.22    129.70  +8.54e+04
1       7       9.12    9.12    129.69  +6.03e+05
6       7       9.12    0.61    129.69  +1.05e+05
3       7       9.12    4.56    129.69  +3.51e+05
5       7       9.12    1.64    129.67  +1.68e+05
```

```
1230     8       9.11    1.40    129.60  +9.81e+04
1229     8       6.69    9.17    129.48  +6.23e+04
7        9       6.71    7.70    129.34  +1.79e+05
11       9       6.70    1.58    129.34  +1.53e+05
10       9       6.70    1.97    129.31  +1.38e+05
9        9       6.70    4.18    129.28  +1.94e+05
8        9       6.69    6.70    129.22  +1.96e+06
15       10      6.69    1.17    129.20  +5.97e+05
```

The frequencies and intensity columns (no requirement for the order of the columns) are used by AutoStructure. Other columns are ignored.

I.4 J coupling list --- in BMRB NMR-STAR 2.1 format

```
#
#      Coupling Atom1   Atom1           Atom2   Atom2
#      Constant Seq     Residue Atom1   Seq     Residue Atom2
#ID    Code     Code    Label   Name    Code    Label   Name    Value   Error
#-------------------------------------------------------------------------------
#
1       3JHNHA  4       SER     H       4       SER     HA      6       1.5
2       3JHNHA  5       SER     H       5       SER     HA      5.4     1.5
3       3JHNHA  6       SER     H       6       SER     HA      4.5     1.5
4       3JHNHA  7       LEU     H       7       LEU     HA      4.6     1.5
5       3JHNHA  8       GLY     H       8       GLY     HA      4.3     1.5
6       3JHNHA  9       ALA     H       9       ALA     HA      4       1.5
7       3JHNHA  10      VAL     H       10      VAL     HA      5.1     1.5
```

I.5 Slow NH exchange list

```
32 LEU
34 CYS
39 PHE
40 PHE
42 ARG
43 ILE
44 HIS
50 ASP
51 GLY
52 VAL
53 ARG
54 GLU
55 LYS
56 SER
57 ASP
59 HIS
60 ILE
61 LYS
62 LEU
63 GLN
64 LEU
67 GLU
```

I.6 Other manual constraints --- in structure calculation software format

- If using Dyana for structure calculations, use the dyana constraints format.
- If using Xplor for structure calculations, use the xplor constraint format.

I.7 Generate Control-file from the interface

Control file has three sections:

1. General Section

- Give the name of the protein, the names of the sequence file, resonance assignment file, J-list file, NH slow exchange file --> all experiment files except peak lists.
- Specify the maximum number of iterative_fold cycles. AutoStructure will stop when no more assignments can be made or after the maximum number of iterative_fold cycles.
- Give the parameter file (optional)

Cis-proline residues are listed as 10, 20, etc. Disulfide bonds are listed as 30-60, 70-100, etc.

2. Command Section - This section specifies how to run HYPER, CreateProc within AutoStructure

All command script are in the $ASBIN directory

CreateProc is used to run structure calculations using Xplor or Dyana.

Options:

```
 -na  Name of the protein
   NOTE: this is mandatory,no default value
-np  Number of processor (queues) to be taken,
           def 10
-ns  Number of structures to be calc. on each processor,
           def  6
-nb  Number of final str. selected,
           def 10
-qu  [PBS|DQS|NOQ] Que system to use, one of PBS or DQS, or NOQUE
    def PBS
```

Examples:

```
for a linux cluster using PBS queneing system:
    $ASBIN/CreateProc -na ER14 -np 14 -nstr 4 -nb 10 -que PBS

for a single cpu:
    $ASBIN/CreateProc -na ER14 -np 14 -nstr 4 -nb 10 -que NOQ
```

When manual analysis results such as upper limit distance constraints, dihedral angle constraints and h-bonds are available; these constraints can be added in and used in structure calculation (optional)

CNS is not supported in this release.

3. PeakList Section - This section gives the name of the peak lists and specifies how to interpret them

For each column in the peak list file, specify the following:		
Attributes	Values	Comments
ColumnType	intensity	the corresponding column is the intensity column
	hx1	the corresponding column is the hx1 frequency column
	hx2	the corresponding column is the hx2 frequency column (for 3D NOESY, it is the independent proton dimension)
	x1	the corresponding column is the x1 frequency column (not used for 2D NOESY) and hx1--> x1
	x2	the corresponding column is the x2 frequency column (only used for 4D NOESY) and hx2-->x2
	label	the corresponding column is the label column (optional)
	id	the corresponding column is the peak id column (optional)
		other types of columns, ignored by the program
AtomType	H	atom type of the corresponding frequency column is H
	N15	atom type of the corresponding frequency column is N15
	C13	atom type of the corresponding frequency column is C13
Tolerance	0.03-0.05(ppm)	recommended tolerance range for H dimension
	0.3-0.5(ppm)	recommended tolerance range for N/C dimension
SweepWidth	0-10000 (ppm)	It is used to determine all possible unaliased chemical shift positions. The program may run faster given a large sweep width for unaliased dimension, such as sw=10000.
Reference	0-10000 (ppm)	It is used for global referencing. If x1.shift=0.1, then all chemical shifts in x1 dimension are added by 0.1ppm. If spectrum is well referenced with resonance assignments, set all =0.
Half-dwell Sampling		If selected, half-dwell sampling is used to filter the possible assignments list. If not selected, no half-dwell sampling filter is applied. Both positive and negative peaks are considered in both cases. Note: two proton dimensions need to have same sweep width (be symmetric). **We assume that positive peaks are folded even times and negative peaks are folded odd times.** If not, set reference = the sweep width to adjust for the phasing problem.

For each peak list:		
Attributes	Values	Comments
have intrachain peaks have interchain peaks		for monomer: select *have intrachian peaks
in d2O solution		

I.8 The original control-file for AutoStructure and RPF

1. General Section:

Keywords:

proteinName	the name of the protein
seqFile	the sequence file in bmrb format
chemicalShiftFile	the resonance assignment file in bmrb format
JListFile	J-List file in bmrb format recommended
NHSlowList	NH slow exchange file (recommended) AutoStructure use slow amide exchange data to determine secondary structures and identify hydrogen bonds.
ACO	dihedral angle constraint file (optional)
HBOND	h-bonds file (optional)
UPL	dyana upper-limit distance constraint file (optional)
LOL	dyana lower-limit distance constraint file (optional)
XPL	xplor distance constraint file for xplor structure calculation optional
par	parameter file (for advanced usage) optional
nCycles	nCycles is the maximum number of iterative-fold cycles (default 10)

Example 1:

```
[General]
proteinName=FGF
#input files except peak lists
seqFile=INPUT/seq.bmrb
chemicalShiftFile=INPUT/chemicalshift.bmrbStereo
JListFile=INPUT/FGFJval.bmrb
NHSlowList=INPUT/NHSlowList
ACO=INPUT/FGF.acoManual
HBOND=INPUT/hbond.dyaManual
UPL=INPUT/FGF.uplManual
```

```
#max. number of iterative_fold cycles
nCycles=10
```

2. Command Section:
Keywords:

hyperCommand	hyper command, all option of hyper can be added at the end of -N (optional) (example)
xplorCommand	CreateProc is a script to run parallel Xplor computing over the DQS/PBS system. `CreatProc -hotad 5000 -hotcy 6 -np 14 -nstr 4 -nb 10 -que PBS' means calculating #np *#nstr (14*4) structures using #np (=14) processors and selecting the best #nb (=10). On each processor, there are #nstr (=4) structures calculated. PBS queueing system is used.  For a single process, use `CreatProc -hotad 5000 -hotcy 6 -np 14 -nstr 4 -nb 10 -que NOQ'
dyanaCommand	Similar description as for xplorCommand
cnsCommand	a script to run cns command (under development, not released in this version)

Example 1:

```
# there can only be one line after each command entry which is treated as a shell
command line and can be commented out.
#here only dyanaCommand is active
# xplorCommand: calc structures on 14 machines, each calc 4 and select best 10
[viewerCommand]
#$ASBIN/View
[hyperCommand]
#$ASBIN/hyper -N
[xplorCommand]
$ASBIN/CreateProc -hotad 5000 -hotcy 6 -np 14 -nstr 4 -nb 10 -que PBS
```

3. PeakList Section: Each peak list is an entry in control file.
Keywords:

dimension	the dimension of peak list dimension = 2 means that the peak list has hx1 and hx2 dimensions. dimension = 3 means that the peak list has hx1, x1 and hx2 dimensions. dimension = 4 means that the peak list has hx1, x1, hx2 and x2 dimensions. Only 4D CC-NOESY is supported right now. Note: hx1--> x1 and hx2--> x2
IC haveIC	for monomer: IC = 0, haveIC = 0
waterFlag	if in water solution, waterFlag = 1; if in D20 solution, waterFlag = 0;

column	col.intensity	the intensity column
	col.label	the label column (optional)
	col.id	the id column (optional)
	col.hx1	the hx1 column
	col.hx2	the hx2 column (for 3D NOESY, it is the independent dimension)
	col.x1	the x1 column (not used for 2D NOESY, if dimension = 2) and hx1--> x1
	col.x2	the x2 column (not used for 2D or 3D NOESY, if dimension = 2 or 3) and hx2-->x2
tol	hx1.tol hx2.tol x1.tol x2.tol	Match tolerance for hx1, hx2, x1 and x2 dimensions in ppm. for H, 0.03-0.05ppm for N/C, 0.3-0.5ppm x1.tol and x2.tol are not used for 2D NOESY x2.tol is not used for 3D NOESY
sw	hx1.sw hx2.sw x1.sw x2.sw	Sweep width for hx1, hx2, x1 and x2 dimensions in ppm. It is used to determined all possible unaliased chemical shift positions. The program may run faster given a large sweep width for unaliased dimension, such as sw=10000.
shift	hx1.shift hx2.shift x1.shift x2.shift	`shift' is used for global referencing. If x1.shift=0.1, then all chemical shift in x1 dimension are added by 0.1ppm. If spectrum is well referenced with resonance assignments, set all shift=0.
type	hx1.type hx2.type x1.type x2.type	atom type for hx1, hx2, x1 and x2 dimensions. type = H for proton type = N15 for nitrogen type = C13 for carbon
sign	hx1.sign hx2.sign x1.sign x2.sign	sign=1 tells the program to use half-dwell sampling to filter the possible assignments list. sign=0 tells the program no half-dwell sampling filter is applied. Both positive and negative peaks are considered in both cases. Note: two proton dimensions need to have same sweep width (be symmetric). We assume that positive peaks are folded even times and negative peaks are folded old times. If not, set shift = the sweepwidth to adjust for the phasing problem.

Example : Peak list entries for a monomer protein (the red-color word need to be changed for your data set)

2D NOESY 3D N15-NOESY 3D C13-NOESY 4D CC-NOESY

```
[INPUT/2d.noesy]
# line above is
the peak list
file name
#it is a 3d
noesy
dimension=2
 IC=0
haveIC=0
#in h2o
waterFlag=1
# intensity is in
column 5
# label is in
column 2
# id is in column
1
# hx1 is in
column 3
# hx2 is in
column 4
col.intensity=5
col.label=2
col.id=1
col.hx1=3
col.hx2=4



# hx1: the match
tolerance, sweep
width, global
reference, sign,
atom type
hx1.tol=0.03
hx1.sw=1000
hx1.shift=0
hx1.sign=0
hx1.type=H
# hx2
hx2.tol=0.05
hx2.sw=1000
hx2.shift=0
hx2.sign=0
hx2.type=H
```

```
[INPUT/n15.noesy]
# line above is
the peak list file
name
#it is a 3d noesy
dimension=3
IC=0
haveIC=0
#in h2o
waterFlag=1
# intensity is in
column 6
# label is in
column 2
# id is in column
1
# hx1 is in column
3
# hx2 is in column
4
# x1 is column 5
col.intensity=6
col.label=2
col.id=1
col.hx1=3
col.hx2=4
col.x1=5


# hx1: the match
tolerance, sweep
width, global
reference, sign,
atom type
hx1.tol=0.05
hx1.sw=13.44
hx1.shift=0
hx1.sign=0
hx1.type=H
# hx2
hx2.tol=0.05
hx2.sw=1000
hx2.shift=0
hx2.sign=0
hx2.type=H
# x1
x1.tol=0.5
x1.sw=27.0
x1.shift=0
x1.sign=0
x1.type=N15
```

```
[INPUT/c13.noesy]
# line above is
the peak list file
name
#it is a 3d noesy
dimension=3
IC=0
haveIC=0
#in d2o
waterFlag=0
# intensity is in
column 6
# label is in
column 2
# id is in column
1
# hx1 is in column
3
# hx2 is in column
4
# x1 is column 5
col.intensity=6
col.label=2
col.id=1
col.hx1=3
col.hx2=4
col.x1=5


# hx1: the match
tolerance, sweep
width, global
reference, sign,
atom type
hx1.tol=0.05
hx1.sw=9.16
hx1.shift=0
hx1.sign=0
hx1.type=H
# hx2
hx2.tol=0.05
hx2.sw=1000
hx2.shift=0
hx2.sign=0
hx2.type=H
# x1 - aliased
x1.tol=0.5
x1.sw=20.7
x1.shift=0
x1.sign=1
x1.type=C13
```

```
[INPUT/c13.noesy]
# line above is
the peak list file
name
#it a 4d noesy
dimension=4
IC=0
haveIC=0
#in d2o
waterFlag=0
# intensity is in
column 7
# label is in
column 2
# id is in column
1
# hx1 is in column
3
# hx2 is in column
4
# x1 is column 5
# x2 is column 6
col.intensity=7
col.label=2
col.id=1
col.hx1=3
col.hx2=4
col.x1=5
col.x2=6
# hx1: the match
tolerance, sweep
width, global
reference, sign,
atom type
hx1.tol=0.05
hx1.sw=9.16
hx1.shift=0
hx1.sign=0
hx1.type=H
# hx2
hx2.tol=0.05
hx2.sw=1000
hx2.shift=0
hx2.sign=0
hx2.type=H
# x1 - aliased
x1.tol=0.5
x1.sw=20.7
x1.shift=0
x1.sign=1
x1.type=C13
# x2 - aliased
x2.tol=0.5
x2.sw=20.7
x2.shift=0
x2.sign=1
x2.type=C13
```

Appendix II Output files

Files under the output directory (output_dir):

*_NA.ovw (***)		General report about AutoStructure/RPF calculation. M-score and global referencing are also reported here. When AutoStructure stops normally, it will print 'The program is finished.' at the end of the file.
*_NA.sec (***)		information about secondary structure analysis
*_NA.exm (***)		Complete report about the initial_fold analysis, providing information about why peaks were assigned or not assigned and the reasoning. It also gives clues for chemical shift refinement.
*_NA.unassign (**)		List of peaks that unassigned during valication cycles.
*_NA.note		report of the preprocessing of the inputfiles and input data errors
*.noise		list of peaks that excluded from noesy_assign analysis – no matches
*_NA.QM		Summary of M scores, calculated using -v. It provides guidance for chemical shift refinement. For atoms that shifts consistently more than 0.3ppm for C/N or 0.03ppm for H, it is recommended to adjust them manually.
*_NA.log		Log file
*_NA.QI		RPF reports summary of Recall scores, and provides list of peaks that are not explained by the input structures
*_NA.QL		RPF reports summary of Precision scores and provides list of back calculated peaks that are missing from the peak lists
source		a place that save all input files used in this calculation, including par.tbl and control-file.
CYCLEs	WorkingCycle	intermideiate DYANA calculation
	CYCLE1-0	initial_fold analysis results
	CYCLE*-0	iterative fold analysis cycles
	Others	Validation cycles
		The final structure and constraints are in the last cycle

Files in each CYCLE: (Before structure calculation cycle is finished, all constraint files and intermediate results are stored in WorkingCycle directory.)

Peak list analysis/assignment results	
*_assOrder (**)	Report NOE assignments for each peaks ordered by intra, seq, mid and long-range.

*_assSparky (**)	report NOE assignments in Sparky format. It can be loaded into Sparky and display assignment in SPARKY. You can convert this file to other format for display.
*_match.gz (*)	report the complete match lists generated by NOESY_match with analysis reasoning records.
RefinedShift.bmrb (**)	Refined chemical shifts used in this cycle.
Xplor input files - manual input constraints are concatenated at the end of the corresponding constraint files	
*.seqXplor	sequence file
*_noe.tbl (***)	distance constraint file
*_dihe.tbl (***)	dihedral angle constraints
*_hbond.tbl (***)	h-bond distance constraints
sa_for_AS*.inp	scripts for xplor calculations
DYANA input files - manual input constraints are concatenated at the end of the corresponding constraint files	
*.seq	sequence file
*.upl (***)	upper limit distance constraints.
*.aco (***)	dihedral angle constraints.
hbond.lol (***)	h-bond lower limit distance constraints.
hbond.upl (***)	h-bond upper limit distance constraints.
Xplor output files	
*.pdb (***)	best 10 XPLOR coordinates with IUPAC naming
*_sa_*.pdb	structures generated from XPLOR
sa_for_AS*.inp.out (*)	xplor outputs for each structure calculations
Violations.*	violation analysis from PDBStat
DYANA output files	
*.pdb (***)	DYANA coordinates with IUPAC naming
*.ovw (*)	DYANA summary file
Log*	DYANA calculation log files for structure caclulation on every machine in the pc-cluster

Others	log or intermediate files for Hyper, parallel calculations, etc