Workshop on High Throughput NMR Structure Determination of Proteins in the Post-Genomic Era

Organized by

High-Field Biomacromolecular Solution NMR Core Facility,

National Research Program for Genomic Medicine, Taiwan, R.O.C.

Major Sponsors

National Science Council, Taiwan, R.O.C.

Taiwan Biophysical Society

Institute of Biomedical Sciences, Academic Sinica, Taiwan, R.O.C.

Corporate Sponsor

Bruker BioSpin GmbH

Contact Information

NMR Core Facility
Institute of Biomedical Sciences, Academia Sinica
No.128, Sec 2, Yen-Chiu-Yuan Rd. 115 Taipei, Taiwan, R.O.C.
(http:// www.nmr.sinica.edu.tw)

Program

Saturday, November 1:

08:00 – 09:00	Registration and Continental Breakfast
09:00 – 09:15	Opening remarks (**Sunney Chan and Andrew Wang**)
Session I	**Chair: Lou-Sing Kan** (B1C Auditorium)
09:15 – 09:25	Overview **(Tai-Huang Huang)**
09:25 – 09:55	The view from the dark side: Dealing with flexible, aggregated and sparingly soluble targets **(Ray Norton)**
09:55 – 10:20	High throughput screening of soluble recombinant proteins **(Tin-Fan Wang)**
10:20 – 10:35	Survey on cell free protein expression **(Shih-Che Sue)**
10:35 – 10:50	Coffee break
Session II	**Chair: Chinpan Chen**
10:50 – 11:40	Isotope filtered/edited NMR methods for the study of biomolecular complexes - acquisition & optimization **(Ruediger Weisemann)**
11:40 – 12:30	Multidimensional NMR spectral processing and experimental scheme for NMR structural genomics **(Guang Zhu)**
12:30 – 13:30	Lunch break
Laboratory Session	**Coordinators : Chi-Fon Chang & Winston Wu** (B1A Conference Room)
13:30 – 14:10	Sample design and manipulations **(Ray Norton)**
14:10 – 15:50	NMR data acquisition and processing **(Ruediger Weisemann)**
15:50 – 17:30	NMR data processing **(Guang Zhu)**

Sunday, November 2:

08:00 – 08:30	Continental Breakfast
Session III	**Chair: Ping-Chiang Lyu** (B1C Auditorium)
08:30 – 09:20	Tracked automated NMR assignments in proteins (TATAPRO) for high throughput structure determination **(K.V.R. Chary)**
09:20 – 10:10	Computer-aided NMR structure determination **(Peter Güntert)**
10:10 – 10:25	Survey on methods for ordering proteins in solution **(Chung-Ke Chang)**
10:25 – 10:40	Coffee break
Session IV	**Chair: Shan-Ho Chou**
10:40 – 11:30	Stereo-Array Isotope-Labeling (SAIL) method: High throughput and accurate structural determinations of proteins **(Masatsune Kainosho)**
11:30 – 12:20	Residual dipolar couplings in the selection and characterization of structural genomics targets **(James Prestegard)**
12:20 – 13:30	Lunch break
Laboratory Session	**Coordinators : Chi-Fon Chang & Winston Wu** (B1A Conference Room)
13:30 – 14:40	TATAPRO-Tracked Automated Assignments in Protein **(K.V.R. Chary)**
14:40 – 16:00	Computer-aided NMR structure determination **(Peter Güntert)**
16:00 – 17:30	REDCAT – Residual Dipolar Coupling Analysis Software Tool **(James Prestegard)**

November 1st

Handout for lecture

(B1C Auditorium)

The view from the dark side: dealing with flexible, aggregated and sparingly soluble targets

Raymond S. Norton

The Walter and Eliza Hall Institute of Medical Research, Parkville, 3050, AUSTRALIA

With access to high-field NMR spectrometers equipped with cryoprobes, solving the solution structures of polypeptides and proteins has become simpler and faster. If the proteins are double-labelled and the full suite of triple resonance 3D spectra can be acquired, this is also true for the tasks of assigning the spectra and calculating the structure. High-throughput structure determination by NMR has become a reality. But there is an important prerequisite, namely a well-behaved protein that is able to be expressed in good yield and thus isotope-labelled, as well as being soluble, stable, and not prone to aggregation. A further complication arises when your favourite protein has significant regions of poorly ordered backbone, making the concept of a single structure less relevant.

One advantage of NMR is that it can tackle proteins that suffer from some or all of these problems and still provide valuable information, even in the absence of a well-defined 3D structure. In this respect it has a significant advantage over X-ray crystallography. My talk will describe our work on domains of malarial surface antigens [1] and on the insulin-like growth factors [2] and their binding proteins, which in many respects are 'badly behaved' in solution. Nonetheless, we have been able to obtain biologically useful information on these proteins, even though our work could in no sense be described as high-throughput structure determination. Some strategies for dealing with badly behaved proteins such as these will be discussed.

1. Nair M, Hinds MG, Coley AM, Hodder AN, Foley M, Anders RF, Norton RS (2002) Structure of domain III of the blood-stage malaria vaccine candidate, *Plasmodium falciparum* apical membrane antigen 1 (AMA1). *J Mol Biol* 322, 741-753.
2. Torres, A.M., Forbes, B.E., Aplin, S.E., Wallace, J.C., Francis, G.L. & NORTON, R.S. (1995) Solution structure of human insulin-like growth factor II. Relationship to receptor and binding protein interactions. *J Mol Biol* 248, 385-401.

High-throughput screening of soluble recombinant proteins

Ting-Fang Wang (tfwang@gate.sinica.edu.tw)

Inst. of Biological Chemistry, Academia Sinica, Taiwan

The aims of high-throughput (HTP) protein production systems are to obtain well-expressed and highly soluble proteins, which are preferred candidates for use in structure–function studies. Here, we describe the development of an efficient and inexpensive method for parallel cloning, induction, and cell lysis to produce multiple fusion proteins in *Escherichia coli* using a 96-well format. Molecular cloning procedures, used in this HTP system, require no restriction digestion of the PCR products. All target genes can be directionally cloned into eight different fusion protein expression vectors using two universal restriction sites and with high efficiency (>95%). To screen for well-expressed soluble fusion protein, total cell lysates of bacteria culture (~1.5 mL) were subjected to high-speed centrifugation in a 96-tube format and analyzed by multiwell denaturing SDS-PAGE. Our results thus far show that 80% of the genes screened show high levels of expression of soluble products in at least one of the eight fusion protein constructs. The method is well suited for automation and is applicable for the production of large numbers of proteins for genome-wide analysis.

1. Yan-Ping Shih, Wen-Mei Kung, Jui-Chuan Chen, Chia-Hui Yeh, Andrew H.-J. Wang and **Ting-Fang Wang** (2002) High throughput screening of soluble recombinant proteins. ***Protein Science*** 11, 1714-1719. (Corresponding author)
2. **Ting-Fang Wang** *and* Andrew H.-J. Wang (2003) High throughput screening of soluble recombinant proteins. Chapter 5 in "*Purifying Proteins for Proteomics: A Laboratory Manual*". Edited by Richard Simpson. **Cold Spring Harbor Laboratory Press**, New York. USA. (Corresponding author)

High-throughput screening of soluble recombinant proteins

YAN-PING SHIH,[1] WEN-MEI KUNG,[1] JUI-CHUAN CHEN, CHIA-HUI YEH, ANDREW H.-J. WANG, AND TING-FANG WANG

Institute of Biological Chemistry, Academia Sinica, Taipei 115, Taiwan, Republic of China

(RECEIVED February 18, 2002; FINAL REVISION April 9, 2002; ACCEPTED April 12, 2002)

Abstract

The aims of high-throughput (HTP) protein production systems are to obtain well-expressed and highly soluble proteins, which are preferred candidates for use in structure–function studies. Here, we describe the development of an efficient and inexpensive method for parallel cloning, induction, and cell lysis to produce multiple fusion proteins in *Escherichia coli* using a 96-well format. Molecular cloning procedures, used in this HTP system, require no restriction digestion of the PCR products. All target genes can be directionally cloned into eight different fusion protein expression vectors using two universal restriction sites and with high efficiency (>95%). To screen for well-expressed soluble fusion protein, total cell lysates of bacteria culture (~1.5 mL) were subjected to high-speed centrifugation in a 96-tube format and analyzed by multiwell denaturing SDS-PAGE. Our results thus far show that 80% of the genes screened show high levels of expression of soluble products in at least one of the eight fusion protein constructs. The method is well suited for automation and is applicable for the production of large numbers of proteins for genome-wide analysis.

Keywords: Structural genomics; functional genomics; proteomics; protein expression

The function of a gene is manifested by the protein it encodes. Genome sequencing of many organisms (see http://www.ncbi.nlm.nih.gov/) has led to the concept of analyzing protein function on a genome-wide scale. Structural genomics and proteomics (Christendat et al. 2000; Skolnick et al. 2000; Fields 2001), therefore, have become major research foci. The challenge of studying proteins in a global scale is driving the development of high-throughput (HTP) and parallel approaches in protein expression, purification, biochemical analysis, and structure determination.

Several prototypes of HTP protein expression and purification systems have been initiated (Christendat et al. 2000; Edwards et al. 2000; Lesley 2001; Zhu et al. 2001). Cloning and expression in *Escherichia coli* are favored in many instances because *E. coli* has relatively simple genetics, is well characterized, has a relatively rapid growth rate, and has few post-translational protein modifications. One disadvantage, however, of expressing heterologous proteins in *E. coli* is that proteins are frequently expressed as insoluble aggregated folding intermediates, known as inclusion bodies (Paul et al. 1983). Although it may be possible to increase protein solubility by optimizing expression condition or by refolding the recombinant proteins, in the interests of throughput, only a single set of growth or folding conditions can be used.

Gene fusion is another approach that has been successfully used for producing soluble heterologous proteins in *E. coli* (Uhl'en and Moks 1990). Several carrier proteins are widely used in gene fusion, including thioredoxin (Trx), maltose-binding protein (MBP), glutathione S-transferase (GST), intein, calmodulin-binding protein (CBP), NusA, and cellulose-associated protein (CAP). Although the use of these carrier proteins has resulted in the successful overexpression of many heterologous proteins, each was tested

Reprint requests to: Ting-Fang Wang, Institute of Biological Chemistry, Academia Sinica, Taipei 115, Taiwan, Republic of China; e-mail: tfwang@gate.sinica.edu.tw; fax: 886-2-27889759.

[1]These two authors contributed equally to this work.

Abbreviations: HTP, high throughput; IPTG, isopropyl β-D-thiogalactoside; LB, Luria-Bertani; RC, recombinational cloning; PCR, polymerase chain reaction; Trx, thioredoxin; MBP, maltose-binding protein; GST, glutathione S-transferase; CBP, calmodulin-binding protein; CAP, cellulose-associated protein; SDS-PAGE, sodium dodecyl sulfate-polyacrylamide gel.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.0205202.

empirically and certainly may not possess maximal solubilizing characteristics. Moreover, each expression scenario requires a specific vector. Recloning cDNA into each of these specific vectors is extremely labor intensive.

Recombinational cloning (RC) methodology was recently developed to minimize the effort required for alternative expression. It uses either cre-lox (Liu et al. 1999) or Int/Xis/IHF (Hartley et al. 2000) recombination to introduce the gene of interest into a recipient vector. In these systems, aberrant recombination or cointegrant products may result from faulty gene transfer to the expression vector. Another limitation is that translation fusions of the recombination *att* or *lox* sites and a few extra nucleotide sequences are required to ensure successful gene transfer. In some cases, such as protein crystallography, in which longer translation fusions are potentially more detrimental to the proteins, a conventional cloning approach with shorter translation fusions is more appropriate. In the present study, we established a new procedure for the parallel cloning of genes into multiple fusion expression vectors without restriction digestion. The main objective here was to rapidly screen for well-expressed soluble proteins that can be used in structural and functional genomics.

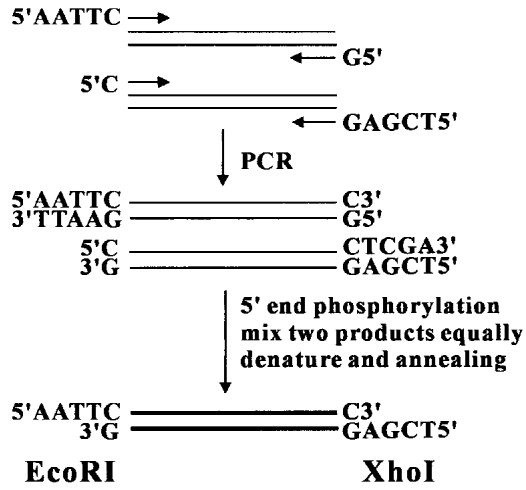


Fig. 1. Moleclular cloning strategy. Four PCR primers and reactions were used in two separate tubes. An equal amount of the two PCR products were mixed, and then the 5′ ends were phosphorylated with T4 polynucleotide kinase. After denaturing (95°C for 5 min) and renaturing (65°C for 10 min), ~25% of the final products carry EcoRI (5′) and XhoI (3′) cohesive ends and are ready for ligation with the vectors.

Results

Parallel cloning of target genes into multiple fusion protein expression vectors

We applied the "sticky end PCR method" (Zeng 1998) to generate DNA products with 5′ EcoRI and 3′ XhoI sticky ends. As illustrated in Figure 1, the method requires four PCR primers and reactions in two separate tubes. Both PCR products were purified and mixed equally. After denaturation and renaturation, ~25% of the final product carries two cohesive ends and is ready for ligation even without restriction digestion. Therefore, this method is suitable for cloning any gene, even genes with internal EcoRI or XhoI restriction sites. To optimize cloning efficiency, sticky-end PCR products were 5′ phosphorylated with T4 polynucleotide kinase and the vectors were dephosphorylated by calf intestine alkaline phosphatase. Together, these procedures increase the efficiency of PCR products into multiple expression vectors. As shown in Figure 2, two independent clones of each ligation reaction were analyzed by restriction digestion. Among the 16 clones of the eight different fusion protein expression vectors, 15 (Fig. 2A) and 16 (Fig. 2B) were identified as successful clones. We applied this method to clone ~40 genes into these eight expression vectors (>300 cloning reactions) with a >95% success rate.

Induction and screening of soluble fusion proteins

Because bacteria host strain JM109(DE3) is suitable for both plasmid DNA preparation and high level protein expression, it was used initially in this investigation. If the digested vectors were tested as efficacious (~100% in a single cloning reaction), bacterial colonies were directly induced with isopropyl β-D-thiogalactoside (IPTG) to produce proteins even without examining each individual clone by restriction mapping or colony PCR.

To identify well-expressed and highly soluble fusion proteins, 2 mL of culture was used for small-scale induction. Briefly, bacterial cultures in log phase (OD_{600}~0.6) were induced with IPTG at 20°C for 24 hr. We found that low

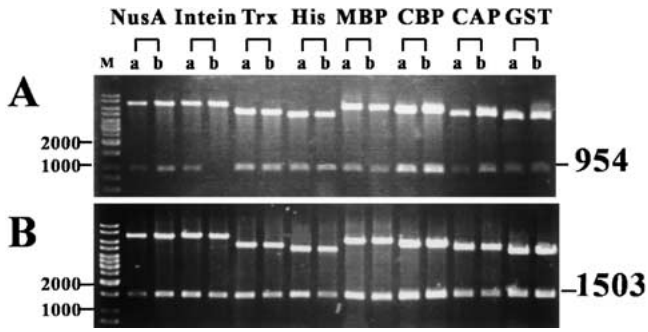


Fig. 2. Recombinant DNA plasmids purified from JM109(DE3). Eight different fusion protein expression vectors are indicated above. Two independent clones from each construct were isolated for characterization (lanes *A* and *B*). Plasmid DNAs were purified in a 96-well format using Millpore's Motage plasmid mini-prep kit; 3–5 μL mini-prep DNA was restriction digested with EcoRI and XhoI and separated in 0.8% agarose gel. A 1-kb DNA ladder (from MBI Fermentas, USA) was used as marker (M) and shown in the far *left* lane. The expected sizes (in base pair) of the desirable restriction fragments of two different target genes are indicated on the *right* of the figure.

temperature and long induction time facilitate correct protein folding; for instance, the fusion protein of yeast Hop2 (encoded by open reading frame YGL033W; Table 1) and Trx is soluble at 20°C but not at 37°C.

In a parallel analysis of protein solubility, host cells were harvested and lysed in 96-well plates as described under Materials and Methods. Insoluble materials in total cell lysates were removed by centrifugation using a Ti25 rotor, which allows parallel processing of 96 samples; therefore, this system is suitable for automation. To increase the accuracy of protein solubility testing, an ultracentrifugal force (90,000*g*) was applied to eliminate partially folded protein aggregates. As illustrated in Figure 3, we applied this HTP system to the expression of yeast Csm2 protein (encoded by open reading frame YIL132C; Table 1). SDS-PAGE was used to separate proteins from total cell lysates induced with or without IPTG induction (Fig. 3, lanes 1 and 2) and from the soluble protein fraction induced with IPTG induction (Fig. 3, lane 3). NusA and MBP fusion proteins were found to be well induced and soluble; on the other hand, GST and Trx fusion proteins were expressed in insoluble forms (Fig. 3).

If the proteins were poorly expressed, the DNA clones were retransformed into other host strains, for example, BL21-Gold(DE3) or BL21-CondonPlus(DE3), in an attempt to alleviate problems related to codon bias or protein toxicity. For example, none of the eight fusion proteins of *Drosophila* Phyl protein (accession number AAF58245; Table 1) were induced in JM109(DE3); on the contrary, NusA-Phyl and GST-Phyl fusion proteins were highly expressed and soluble in BL21-CondonPlus(DE3) (data not shown).

We have cloned and expressed more than 40 proteins from various organisms. The overall successful rate of obtaining soluble proteins, at least in one of the eight expression constructs tested, is >80% (Fig. 4A). The soluble ratio of individual fusion protein is shown in Figure 4B. Often, the larger fusion tags are superior for enhancing protein solubility; for example, the success ratio of soluble NusA (54 kD), MBP (42 kD), and GST (24 kD) fusion proteins are 60%, 60%, and 38%, respectively. Target fusion proteins that have been successfully expressed by this method are listed in Table 1. These target proteins alone range from 9 kD to 100 kD, and the largest soluble fusion protein expressed in this study was ~150 kD.

Table 1. *Soluble target proteins with high expression levels*

Organism	Gene	Protein size (kD)
Yeast	YHL024W	80.1
	YOR351C	56.9
	YLR394W	53.9
	YBR233W	45.8
	YDR065W	42.9
	YPL018W	42.8
	YOL104C	40.9
	YER106W	35.8
	YHR014W	33.3
	YIL144W[a]	28.2
	YPL199C	26.8
	YGL033W	25.0
	YCR086W	21.7
	YIL132C	25.0
	YMR048W	36.3
	YPL200W	18.3
Mammalian	U47110[b]	100.0
	U47110[b]	35.2
	U47110[b]	24.2
	U47110[b]	8.8
	P97801	32.3
	AAC25954	31.0
	NP_064587	30.5
	AJ404613	27.5
	NP_036520	19.7
	XP_043137	16.3
Plant	AAF75761	27.9
	CAC17699[c]	28.4
	CAC17699[c]	28.4
	O04701	26.8
Insect	BAB17671	69.6
	AAF58245	44.0

[a] Only the N-terminal 256 aa was expressed.
[b] Full-length and three truncated proteins were expressed.
[c] Wild type and mutant protein with mutation in the amino acid residue 128.

Generalized protein purification strategy

In an HTP process, it is absolutely essential that purification does not depend on the tedious optimization of conditions that exploit subtle differences in protein size, charge, or hydrophobicity. Therefore, it is advantageous to use expression vectors with multiple tagging for affinity purification. Almost all expression vectors used in this study were engineered with an NH_2-terminal affinity tag, a cleavage site of protease (e.g., thrombin or factor Xa), and a COOH-terminal His-tag. Recombinant fusion proteins were first isolated by various affinity chromatography columns (glutathione agarose, amylose resin, etc) and then further purified by Ni^{2+}-resin. Routinely, fusion proteins with typical yields (5–20 mg per liter of Luria-Bertani [LB] culture) and purity (>90%) have been obtained (Fig. 5). Because all these fusion constructs can be proteolytically cleaved to remove the NH2-terminal fusion partners, it is of interest to examine if the cleaved target proteins are still soluble. Thus far, we have tested three yeast target proteins (Trx-YGL033W, MBP-YPL199C, and Nus-YIL144W; Table 1) with thrombin, and all yielded soluble products (data not shown).

Discussion

In the postgenomic era, HTP protein expression technologies are essential tools. Conceivably, the most important

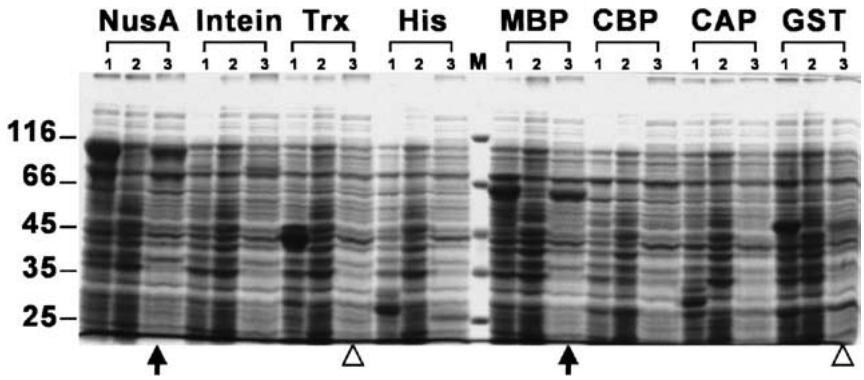


Fig. 3. Analysis of budding yeast Csm2 (YIL132C) fusion proteins. Samples of total proteins and soluble protein fractions were separated on a 10% SDS-PAGE under reducing conditions and stained with Coomassie Blue. Lane *1*, whole cell lysates of induced cells; lane *2*, whole cell lysates of uninduced cells; lane *3*, soluble proteins with induction. Eight different fusion proteins are indicated above. The molecular weight standards are shown in the center and labeled on the *left* (×1,000). NusA and MBP fusion proteins show high solubility (indicated by arrows below the lanes of soluble protein fractions); on the other hand, GST and Trx fusion proteins are well induced but not soluble (indicated by open triangles).

characteristic of a protein that determines its feasibility for functional analysis is its solubility. High solubility is also strongly correlated with the success of structural studies using either NMR or X-ray crystallography. The method described here allows one to clone and express multiple fusion proteins in *E. coli* efficiently. The success ratio for obtaining highly expressed and soluble products in one of the eight fusion constructs is >80%, which is superior to the results of other structural or functional genomic studies (Christendat et al. 2000; Edwards et al. 2000).

Sticky-end PCR and directional cloning methods allow one to obtain multiple expression plasmids without restriction digestion. This is a somewhat conventional cloning approach, but it has several advantages compared with other methods, such as the RC approach. First of all, it is simpler. It allows direct cloning of PCR products into multiple expression vectors. RC methods require at least two cloning steps. Second, it is more accurate in theory and also in practice. With an RC approach, faulty gene transfer might occur because of aberrant recombination or cointegrant vector products. Third, it is less detrimental to proteins. This method introduces only two new amino acids encoded by the restriction sites, whereas RC methods include *att* or *lox* sites, as well as other extra sequences to achieve a precise gene transfer. Longer translation fusions introduced by a cloning procedure are usually more harmful to proteins.

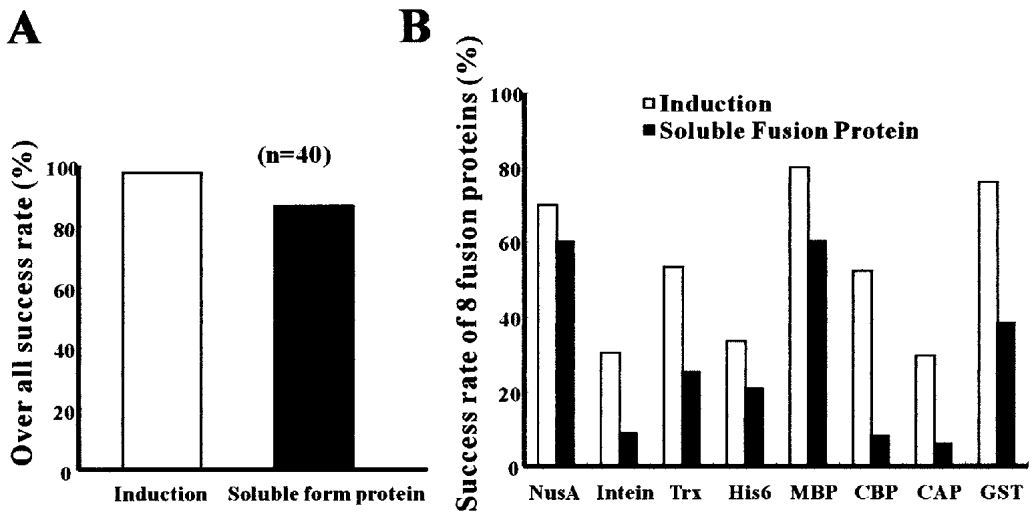


Fig. 4. (*A*) Statistical analysis of soluble protein ratio obtained in at least one of the eight expression constructs. (*B*) Eight different gene fusions and their effects were also compared. A total of 40 different genes were tested in this study. Well-induced and highly soluble fusion proteins were identified visually by comparing the relative density of protein bands in SDS-PAGE as shown in Figure 3.

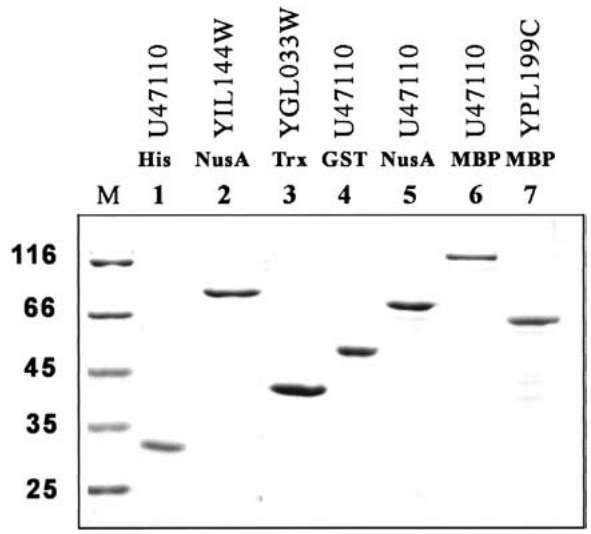


Fig. 5. SDS-PAGE analysis of purified proteins. The gel shows typical yields (5–20 mg per liter of Luria-Bertani [LB] culture) and purity (~90%) obtained from two steps of affinity purification. The database accession or open reading frame number of the expressed proteins and their fusion tags are indicated. The molecular weight standards are labeled on the *left*.

All procedures in this study, including DNA cloning, plasmid preparation, protein induction, and cell lysis are based on using a standard 96-well format in an efficient and reproducible manner, making these procedures suitable for automation. All 96 samples could be subjected to ultracentrifugation fractionation in the Beckman Ti25 rotor. Therefore, it is possible to integrate the process of sample transfer from 96-well plate to 96 centrifugal tubes using a robotic protocol, although manual operation is still required afterward.

High speed centrifugation (90,000*g*) ensures the separation of highly soluble and properly folded proteins from insoluble or partially aggregated materials. Denaturing SDS-PAGE was applied to identify these soluble fusion proteins. Consequently, there is little chance of finding false positives in this screening procedure. After identifying clones expressing soluble fusion proteins, the rest of the cell lysates can be used for other purposes such as small-scale affinity purification.

With our protocols for rapid subcloning, solubility screening, and parallel protein purification, we will be able to provide a large number of high purity fusion proteins for structure–function studies. For protein crystallography, carrier fusion protein domains can be proteolytically cleaved during or after the first step of affinity purification. The resulting His-tagged target proteins can be isolated by Ni^{2+}-resin or other conventional chromatography methods. These proteins can also be used to make protein microarrays, allowing for the parallel characterization of diverse biochemical activities, such as enzymatic assays, protein–protein, protein–nucleic acid, and receptor–ligand interactions. The protein chips may also be applied to screen for new drugs. We have succeeded in the biochemical characterization of at least three fusion proteins expressed in this study (data not shown), indicating that these fusion proteins retain a part of or even the full biochemical activity of the target proteins.

In summary, we have developed an HTP molecular cloning and protein expression system using *E. coli.* It allows us to screen effectively for well-expressed and highly soluble proteins. The same approach can be applied for alternate cloning of all potential target genes into vectors of different expression systems, including yeast, insect, and mammalian cells, as well as cell-free in vitro systems. Last, but not least, this method is well suited for automation and will be a useful tool for the production of proteins for use in structural and functional genomic studies.

Materials and methods

Molecular cloning

A PCR cloning strategy, referred to as the sticky-end PCR method (Zeng 1998), was applied to generate PCR products bearing cohesive ends compatible with 5′ EcoRI and 3′ XhoI sites (Fig. 1). The method requires four PCR primers and reactions in two separate tubes. Both PCR products were purified and mixed equally and then treated with T4 polynucleotide kinase (New England Biolabs) and ATP (Sigma). After denaturing (95°C for 5 min) and renaturing (65°C for 10 min), ~25% of the final products carried cohesive ends and were ready for ligation.

Fusion protein expression vectors used in these studies were purchased from Novagen, New England Biolabs, or Amershan Pharmacia. We engineered two new universal cloning sites (EcoRI and XhoI) into those vectors. Briefly, the original vectors were cut with restriction enzymes as close to the 3′ end of the N-terminal fusion genes. The appropriate DNA cassette was chosen to retain the reading frame of the fusion over EcoRI and XhoI restriction sites and to introduce 6 histidine amino acid residues between XhoI and stop codon. A specific cleavage sequence of protease (e.g., thrombin or factor Xa) was introduced immediately after the EcoRI site and before the coding sequence of target protein; this was achieved by stringent design of the sticky-end PCR primers. To prepare vectors for ligation reactions, the vectors were restriction digested with EcoRI and XhoI and then dephosphorylated with calf intestinal alkaline phosphotase (New England Biolabs).

Plasmid DNA purification was performed in a 96-well format using Millpore's Motage plasmid miniprep kit. Eight different expression vectors were used here for parallel cloning. Two independent clones were isolated and characterized from every cloning reaction. Therefore, soluble protein products of six different genes (48 cloning/96 protein induction) were screened simultaneously.

Small-scale protein induction

Host *E. coli* strain JM109(DE3) (Novagen) was chosen for plasmid preparation as well as protein induction. Host strains, BL21-Gold(DE3) or BL21-CondonPlus(DE3) (Stratagene), were also used for expression in the case of low-level protein induction in JM109(DE3). Single colonies were grown overnight in LB medium with ampicillin (50 μg/mL) or kanamycin (30 μg/mL) at 37°C. Two 18-μL overnight cultures were inoculated in 2 mL LB

(with 1% glucose) and grown at 37°C for 3 hr (OD_{600}~ 0.6). The cells were cooled in 20°C incubators, induced with or without 0.4 mM IPTG, and subsequently grown for an additional 20 hr. To harvest the cells, 500-μL cultures from each well of the 96-wells were transferred to a new 96-well plate. Culture medium was placed onto a Sorvall RTH750 microplate carrier and centrifuged for 10 min at 4000 rpm. Cell pellets were suspended in 1.5X SDS-PAGE sample buffer and boiled for 5 min.

Protein solubility test

For protein solubility assays, cell pellets from 1.5 mL of culture with IPTG induction were resuspended in 40 μL of ice-cold buffer B (250 mM sucrose, 25 mM Tris-HCl at pH 7.0, 1 mM EGTA, lysozyme 0.3 mg/mL) and incubated on ice for 20 min. The suspensions were mixed with 160 μL of ice-cold lysis buffer (0.1% Triton X-100, 150 mM NaCl, 0.1 unit Benzonase, 1 mM EGTA, 25 mM Tris-HCl at pH 7.0) followed by incubation at 4°C for another 20 min. Benzonase (Novagen, USA) was used here to digest bacteria genomic DNA and RNA. Insoluble materials were removed by centrifugation at 90,000*g* for 45 min in the Ti25 rotor (Beckman, USA). Soluble fractions (~100 μL) were then mixed with an equal volume of 3X SDS-sample buffer and boiled immediately for 5 min. Both total cell extracts and soluble fractions were analyzed on 8% to 12% denaturing SDS-PAGE. The proteins (gels) were visualized by Coomassie Blue staining.

Successful expression of soluble fusion protein was scored as follows: Eight different fusion constructs for each target protein were examined. At least one of these constructs must yield a high level of expression and also remain soluble after an ultracentrifugation fractionation procedure, as described previously. Successfully expressed soluble proteins were analyzed by SDS-PAGE and visually identified by Coomassie Blue staining.

Acknowledgments

We gratefully thank Dr. Yi-Ping Hsueh, Su-Ming Hu, and Dr. Chih-Hsiang Leng for helpful discussions; Yu-Jing Hsiao and Shu-Chun Chang for assistance in this study; and Dr. Chung Wang for providing comments on the manuscript. This work was supported by Academia Sinica and National Science Council (NSC90–2321-B-001–015 to A.H.-J. Wang and NSC90–2321-B-001–014 to T.-F. Wang).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J.R., Booth, V., Mackereth, C.D., Saridakis, V., Ekiel, I., et al. 2000. Structural proteomics of an archaeon. *Nat. Struct. Biol.* **7:** 903–909.

Edwards, A.M. Arrowsmith, C.H. Christendat, D. Dharamsi, A., Friesen, J.D. Greenblatt, J.F., and Vedadi, M. 2000. Protein production: Feeding the crystallographers, and NMR spectroscopies. *Nat. Struct. Biol.* [*Suppl.*] **7:** 970–972.

Fields, S. 2001. Proteomics: Proteomics in genomeland. *Science* **291:** 1221–1224.

Hartley, J.L., Temple, G.F., and Brasch, M.A. 2000. DNA cloning using in vitro site-specific recombination. *Genome Res.* **10:** 1788–1795.

Lesley, S.A. 2001. High-throughput proteomics: Protein expression and purification in the postgenomic world. *Protein Expr. Purif.* **22:** 159–164.

Liu, Q., Li, M.Z., Leibham, D., Cortez, D., and Elledge, S.J. 1999. The univector plasmid-fusion system, a method for rapid construction of recombinant DNA without restriction enzymes. *Curr. Biol.* **8:**1300–1309.

Paul, D.C., Van Frank, R.M., Muth, W.L., Ross J.W., and Williams, D.C. 1983. Immunocytochemical demonstration of human proinsulin chimeric polypeptide within cytoplasmic inclusion bodies of *Escherichia coli. Eur. J. Cell Biol.* **31:** 171–174.

Skolnick, J., Fetrow, J.S., and Kolinski, A. 2000. Structural genomics and its importance for gene function analysis. *Nat. Biotechnol.* **18:**283–287.

Uhl'en, M. and Moks, T. 1990. Gene fusion for purposes of expression: An introduction. *Methods Enzymol.* **185:**129–143.

Zeng, G. 1998. Sticky-end PCR: New method for subcloning. *Biotechniques* **25:** 206–208.

Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamzyor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., et al. 2001. Global analysis of protein activities using proteome chips. *Science* **293:** 2101–2105.

Developing Bacteria *In vitro* protein expression system to aid NMR structural determination

Shih-Che Sue

Institute of Biomedical Sciences, Academia Sinica

For high-throughput protein structural determination in modern NMR, it will be indispensable to develop a rapid and reliable protein over-expression system. Recently, *in vitro* protein synthesis in cell-free system has become an important tool. Comparing to *in vivo* protein expression systems the *in vitro* expression systems have several advantages for expressing target proteins that are toxic to the host, proteins that form inclusion bodies or precipitates, or proteins that can be degraded rapidly. Current application of *in vitro* expression systems include: rapid screening of gene products, mutagenesis studies, protein folding process, and incorporation of unnatural or expensive amino acids. In practice, few cell-free systems have been developed and the most frequently used system is based on *Escherichia coli* extract. In comparison to eukaryotic systems, the *E. coli* extract has a relatively simple translational apparatus, thus making it a very efficient protein synthesis system. However, due to rapid degradation of the exogenous RNA through endogenous nucleases, mRNA cannot be used directly as the starting genetic material for protein translation in bacterial extracts. Thus, a one-pot synthesis using a coupled transcription and translation expression system using DNA as template and exogenous T7 RNA polymerase for DNA transcription has been developed. Combining with the recently proposed creatine phosphate/creatine kinase energy regeneration system milligrams quantities of proteins per milliliter of cell extract has been achieved. Since *in vitro* expression system uses amino acids as nutrients for protein expressing, efficient incorporation of specifically isotope-labeled amino acids can be incorporated to produce various labeled proteins. A survey of the published results of other *in vitro* expression systems, such as wheat germ system and reticulocyte system will be presented.

Cell-free production and stable-isotope labeling of milligram quantities of proteins

Takanori Kigawa[1,a], Takashi Yabuki[a,b], Yasuhiko Yoshida[a,c], Michio Tsutsui[c], Yutaka Ito[d], Takehiko Shibata[d], Shigeyuki Yokoyama[a,b,*]

[a] *Cellular Signaling Laboratory, The Institute of Physical and Chemical Research (RIKEN), 2-1 Hirosawa, Wako, Saitama 351-0198, Japan*
[b] *Department of Biophysics and Biochemistry, Graduate School of Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*
[c] *Central Laboratory, Amersham Japan, 2802-1 Hiratsuka, Shiroi, Inba, Chiba 270-1402, Japan*
[d] *Cellular and Molecular Biology Laboratory, The Institute of Physical and Chemical Research (RIKEN), 2-1 Hirosawa, Wako, Saitama 351-0198, Japan*

Received 22 October 1998; received in revised form 30 November 1998

Abstract **We have improved the productivity of an *Escherichia coli* cell-free protein synthesis system. First, creatine phosphate and creatine kinase were used as the energy source regeneration system, and the other components of the reaction mixture were optimized. Second, the *E. coli* S30 cell extract was condensed by dialysis against a polyethylene glycol solution to increase the rate of synthesis. Third, during the protein synthesis, the reaction mixture was dialyzed against a low-molecular-weight substrate solution to prolong the reaction. Thus, the yield of chloramphenicol acetyltransferase was raised to 6 mg/ml of reaction mixture. Stable-isotope labeling of a protein with $^{13}C/^{15}N$-labeled amino acids for NMR spectroscopy was achieved by this method.**

Key words: Cell-free protein synthesis; In vitro translation; Chloramphenicol acetyltransferase; Ras; Dialysis; Nuclear magnetic resonance

1. Introduction

Biochemical and biophysical studies of proteins usually require large-scale preparations of the proteins of interest. A number of unnatural amino acids have been incorporated site-specifically into proteins by the chemical acylation method [1]. Moreover, specifically stable-isotope-labeled proteins for Fourier transform infrared [2] and NMR [3,4] analyses can be produced. It is also expected that challenging proteins, such those prone to aggregation or with toxic properties, can be expressed by a cell-free system. Despite these merits, the low productivity of cell-free protein synthesis systems has limited their application. Many attempts to improve the productivity have been made, mainly for the *Escherichia coli* and wheat germ systems. First, the duration of the reaction has been increased by substrate supplementation during protein synthesis, by methods such as the continuous-flow method [5]. Recently, a semi-continuous flow system using a dialysis chamber [6] and a conventional dialysis system using a disposable dialyzer [7] have been reported. Second, the rate of protein synthesis has been increased by the use of condensed cell extract. The condensation was achieved by ultrafiltration [8,9] or polyethylene glycol (PEG) precipitation [10]. In addition to these improvements, optimization of the reaction conditions was reported [9,11,12]. Thus, it has been established that about 1 mg of protein can be produced per ml of *E. coli* cell-free reaction mixture [6].

In this study, we have improved the productivity of the *E. coli* coupled transcription-translation cell-free protein synthesis system. First, the energy source regeneration system was changed, and the other components were optimized. Second, the *E. coli* cell extract was condensed by dialysis against a PEG-containing solution. Third, this improved cell-free system was applied to protein synthesis with a disposable dialyzer [7]. As a result, about 6 mg of chloramphenicol acetyltransferase (CAT) protein was synthesized per ml of the reaction mixture in 21 h. Moreover, we applied the improved cell-free system with dialysis for the production of a $^{13}C/^{15}N$-labeled Ras protein for NMR spectroscopy, by using a labeled algal amino acid mixture, and successfully measured the HSQC spectrum.

2. Materials and methods

2.1. Template DNA for cell-free protein synthesis

Plasmids pK7-CAT [9] and pK7-Ras [3], which have the T7 promoter and the gene for the CAT protein and the human c-Ha-Ras protein, respectively, were used as the DNA templates. The Ras protein used in this study consisted of 171 amino acid residues, and lacked the C-terminal 18 amino acid residues, which is a better NMR sample than the full-length form [13–16]. The truncated Ras protein has been shown to have the same guanine-nucleotide binding and GTPase activities, and the same NMR chemical shifts and nuclear Overhauser effects for the corresponding residues, as the full-length Ras protein [13,14].

2.2. Reaction conditions for the batch system

The *E. coli* S30 cell extract used for the cell-free protein synthesis was prepared according to Pratt [17] from *E. coli* strain A19 (*metB*, *rna*). The T7 RNA polymerase was prepared according to Zawadzki and Gross [18]. Acetyl phosphate (AP) was purchased from Kohjin, and acetyl kinase (AK) was from Boehringer-Mannheim. The system used as the starting point of our study (the 'initial' system [3]) consisted of (per 15 μl) 55 mM HEPES-KOH (pH 7.5), 1.7 mM DTT, 1.2 mM ATP, 0.8 mM each of CTP, GTP, and UTP, 27 mM phospho*enol*pyruvate (PEP) (Boehringer-Mannheim), 2.0% polyethylene glycol (PEG) 8000 (Sigma), 0.64 mM 3′,5′-cyclic AMP, 68 μM L(−)-5-formyl-5,6,7,8-tetrahydrofolic acid, 175 μg/ml *E. coli* total tRNA (Boehringer-Mannheim), 210 mM potassium glutamate, 27.5 mM ammonium acetate, 13.3 mM magnesium acetate, 0.46 mM L-[^{14}C]leucine (267 MBq/mmol, Amersham), 0.5 mM of each of the

*Corresponding author. Fax: (81) (48) 462-4675.
E-mail: yokoyama@y-sun.biochem.s.u-tokyo.ac.jp

[1] The first two authors contributed equally to this work.

Abbreviations: AK, acetyl kinase; AP, acetyl phosphate; CAT, chloramphenicol acetyltransferase; CK, creatine kinase; CP, creatine phosphate; DTT, dithiothreitol; MWCO, molecular weight cut off; PEG, polyethylene glycol; PEP, phospho*enol*pyruvate; PK, pyruvate kinase

0014-5793/99/$19.00
PII: S0014-5793(98)01620-2

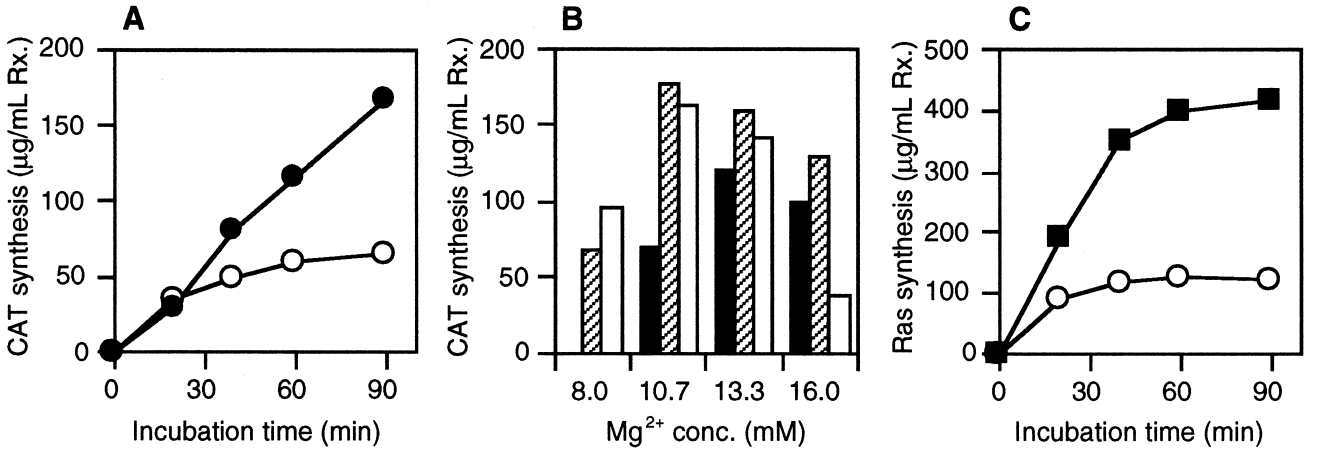


Fig. 1. Improvements of the *E. coli* cell-free system. A: Time courses of CAT synthesis using the system with PEP-PK (○) and the system with CP-CK (●). B: Dependence of CAT synthesis on PEG 8000 concentration (filled bars, 2.0%; hatched bars, 4.0%; open bars, 6.0%) at different magnesium ion concentrations. The incubation time was 1 h. C: Time courses of Ras synthesis by the 'initial' (○) and the 'improved' (■) cell-free systems.

other 19 amino acids, 6.7 μg/ml of either the pK7-Ras plasmid for Ras expression or the pK7-CAT plasmid for CAT expression, 93 μg/ml T7 RNA polymerase, and 3.6 μl S30 extract [3]. On the other hand, the 'improved' system consisted of (per 15 μl) 55 mM HEPES-KOH (pH 7.5), 1.7 mM DTT, 1.2 mM ATP, 0.8 mM each of CTP, GTP, and UTP, 80 mM creatine phosphate (CP) (Boehringer-Mannheim), 250 μg/ml creatine kinase (CK) (Boehringer-Mannheim), 4.0% PEG 8000, 0.64 mM 3′,5′-cyclic AMP, 68 μM L(−)-5-formyl-5,6,7,8-tetrahydrofolic acid, 175 μg/ml *E. coli* total tRNA, 210 mM potassium glutamate, 27.5 mM ammonium acetate, 10.7 mM magnesium acetate, 0.64 mM L-[^{14}C]leucine (193 MBq/mmol, Amersham), 1.0 mM of each of the other 19 amino acids, 6.7 μg/ml of either the pK7-Ras plasmid for Ras expression or the pK7-CAT plasmid for CAT expression, 93 μg/ml T7 RNA polymerase, and 4.5 μl S30 extract. The reaction mixture was incubated at 37°C for 1 h.

2.3. Condensation of S30 extract

The S30 cell extract was placed in a dialysis tube (Spectra/Por, molecular weight cut off (MWCO) 12 000–14 000), and was dialyzed against 10 volumes of an equal-weight mixture of PEG 8000 (Sigma) and the S30 dialysis buffer (10 mM Tris-acetate (pH 8.2), 14 mM magnesium acetate, 60 mM potassium acetate, and 1 mM DTT) in a Heat Seal Bag (Yamamoto) for 45 min at 4°C on a rotator. Then, the S30 extract was dialyzed against 100 volumes of the S30 dialysis buffer for 15 min at 4°C. Typically, the protein was condensed 2–2.5-fold.

2.4. Reaction conditions for the dialysis system

The reaction unit was constructed according to Davis et al. [7] using a DispoDialyzerCE (1 ml, MWCO 10 000 or 50 000, Spectra/Por). The internal solution (300 μl) consisted of the same components used for the improved batch system, 0.05% sodium azide, and 0.5 U/μl RNase Inhibitor (Toyobo). The external solution (3 ml) contained the components of the internal solution except for the creatine kinase, the plasmid vector, the T7 RNA polymerase, the S30 extract, and the RNase inhibitor, and also contained an additional 4.2 mM magnesium acetate corresponding to the magnesium carry over from the S30 extract. The reaction unit was incubated at 37°C with shaking at 160 rpm.

2.5. Assay for reaction products

The incorporation of L-[^{14}C]leucine into the Ras protein was determined by liquid scintillation counting of the trichloroacetic acid-insoluble material. The amount of CAT protein was determined by a colorimetric assay as described [19]. The reaction products were also analyzed by a modified SDS-PAGE [20,21].

2.6. Synthesis of $^{13}C/^{15}N$-labeled Ras protein and NMR analysis

The reaction was carried out using three units of the dialysis system (each unit contained 500 μl of the internal solution and 5 ml of the external solution) with the condensed S30 extract. The amino acids in the internal and external solutions were replaced by 3 mg/ml of the

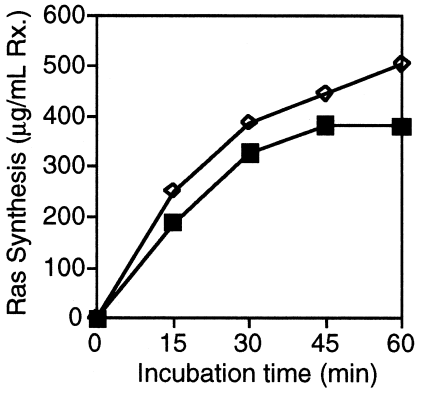


Fig. 2. Time courses of Ras synthesis by the 'improved' cell-free systems with original S30 cell extract (■) and the condensed S30 cell extract (◇).

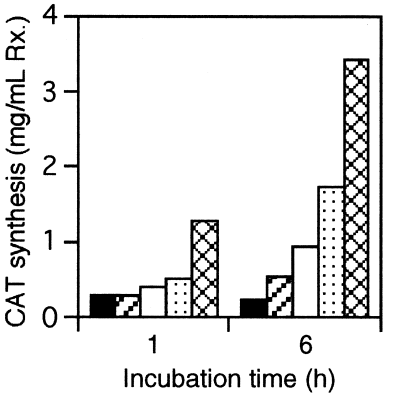


Fig. 3. Dependence of CAT synthesis on the concentration of the S30 extract and the MWCO of the dialysis membrane (filled bars, batch system with original S30; hatched bars, MWCO 10 kDa with original S30; open bars, MWCO 10 kDa with condensed S30; dotted bars, MWCO 50 kDa with original S30; checked bars, MWCO 50 kDa with condensed S30).

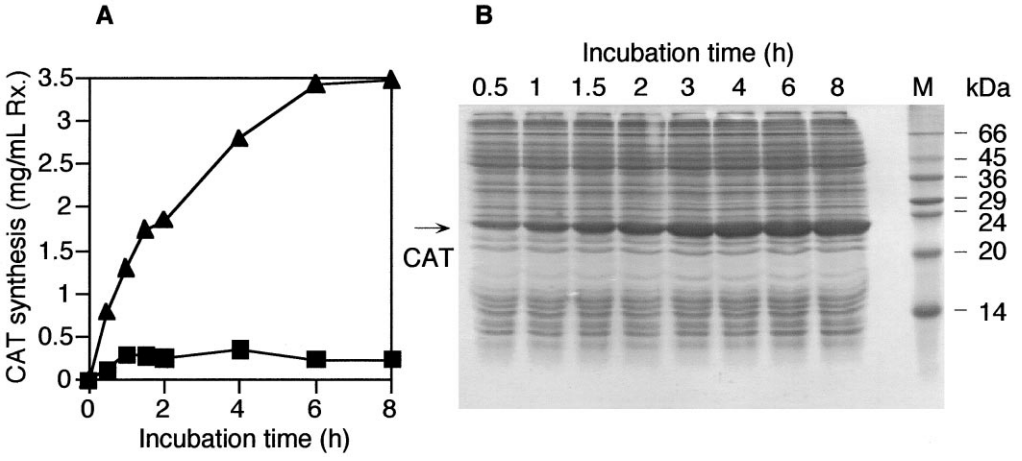


Fig. 4. A: Time courses of CAT synthesis by the 'improved' cell-free batch system (■) and the dialysis system with the condensed S30 cell extract (▲). B: SDS-PAGE analysis of the reaction products of the dialysis system with the condensed S30 cell extract (modified CBB staining [23]).

$^{13}C/^{15}N$-labeled algal amino acid mixture (Chlorella Industries Inc.) supplemented with 1 mM each of L-[^{13}C,^{15}N]phenylalanine and L-[^{13}C,^{15}N]arginine, and with 1 mM each of L-cysteine, L-glutamine, and L-asparagine. The reaction tubes were incubated at 37°C for 4 h. For comparison, the Ras protein was also synthesized in vivo in *E. coli* cells as described [16]. The two samples of the Ras protein were purified and sampled for NMR spectroscopy, and the HSQC spectra were recorded as described [3] except for the pH value (=6.5 in the present study). As the recombinant Ras protein synthesized in vivo retains the N-terminal Met residue, the present Ras protein synthesized in the cell-free system was shown to have the terminal Met residue by Edman sequencing (data not shown).

3. Results and discussion

3.1. Improvement of the productivity of the E. coli cell-free batch system

During the experiments to optimize the 'initial' system, we found that PEP, which was used to regenerate the energy sources, such as ATP and GTP, in the *E. coli* cell-free system, and/or a derivative of PEP, had an inhibitory effect on the system (data not shown). Therefore, instead of using the PEP-PK energy source regenerating system, we examined the AP-AK energy source regenerating system, which enhanced the cell-free protein synthesis over that with PEP-PK [11]. The system with AP-AK (40 mM and 870 μg/ml, respectively, were optimal) was twice as productive as the system with PEP-PK for CAT expression (data not shown). The CP-CK system, which is usually used in eukaryotic cell-free systems [22], was also examined. The cell-free system with CP-CK (80 mM and 250 μg/ml, respectively) gave maximum productivity, and was about 2.5-fold more productive than the system with PEP-PK for CAT expression (Fig. 1A). As the cell-free system with CP-CK was more productive than the system with AP-AK, CP-CK was used in the 'improved' system.

As previously demonstrated by our group [9], the optimal concentrations of PEG and magnesium ion were interdependent. Thus, these concentrations were co-optimized (4.0% PEG and 10.7 mM magnesium acetate were optimal), which increased the CAT expression by about 1.5-fold (Fig. 1B). Furthermore, the volume of the extract used for the cell-free reaction was increased from 3.6 μl/15 μl reaction to 4.5 μl/15 μl reaction, and the concentration of the amino acids was increased from 0.5 mM to 1 mM.

Finally, we compared the system with the newly developed conditions (the 'improved' system, see Section 2) to the system with the previously developed conditions (the 'initial' system) [3] for Ras expression. The 'improved' system was approximately 4-fold more productive than the 'initial' system, and could synthesize more than 0.4 mg of the Ras protein per ml reaction mixture (Fig. 1C).

3.2. Condensation of S30 extract and dialysis system

In our system, condensation with PEG precipitation [8] did not work well, because of the difficulty in dissolving the precipitate, while the reproducibility of condensation with ultrafiltration was not so amenable to being scaled up because of membrane clogging. Therefore, we adopted condensation by dialysis against a PEG-containing solution. This method is simple and can easily be scaled up. The condensed S30 extract appreciably increased both the initial rate of protein synthesis and the total amount of synthesized Ras protein (Fig. 2).

3.3. Dialysis system

By the use of the dialysis system, the amount of synthesized CAT protein was dramatically increased, and the productivity of the dialysis system was dependent on the concentration of the S30 extract and the MWCO of the dialysis membrane

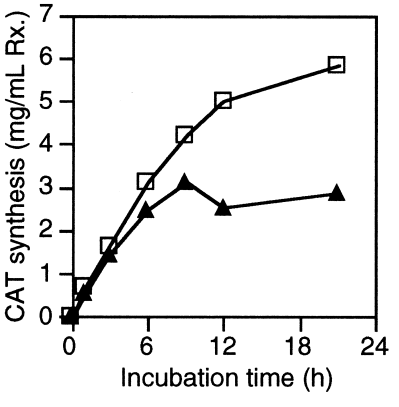


Fig. 5. Time courses of CAT synthesis by the dialysis system using the condensed S30 extract without an exchange of the external solution (▲) and with an exchange at 6 h (□).

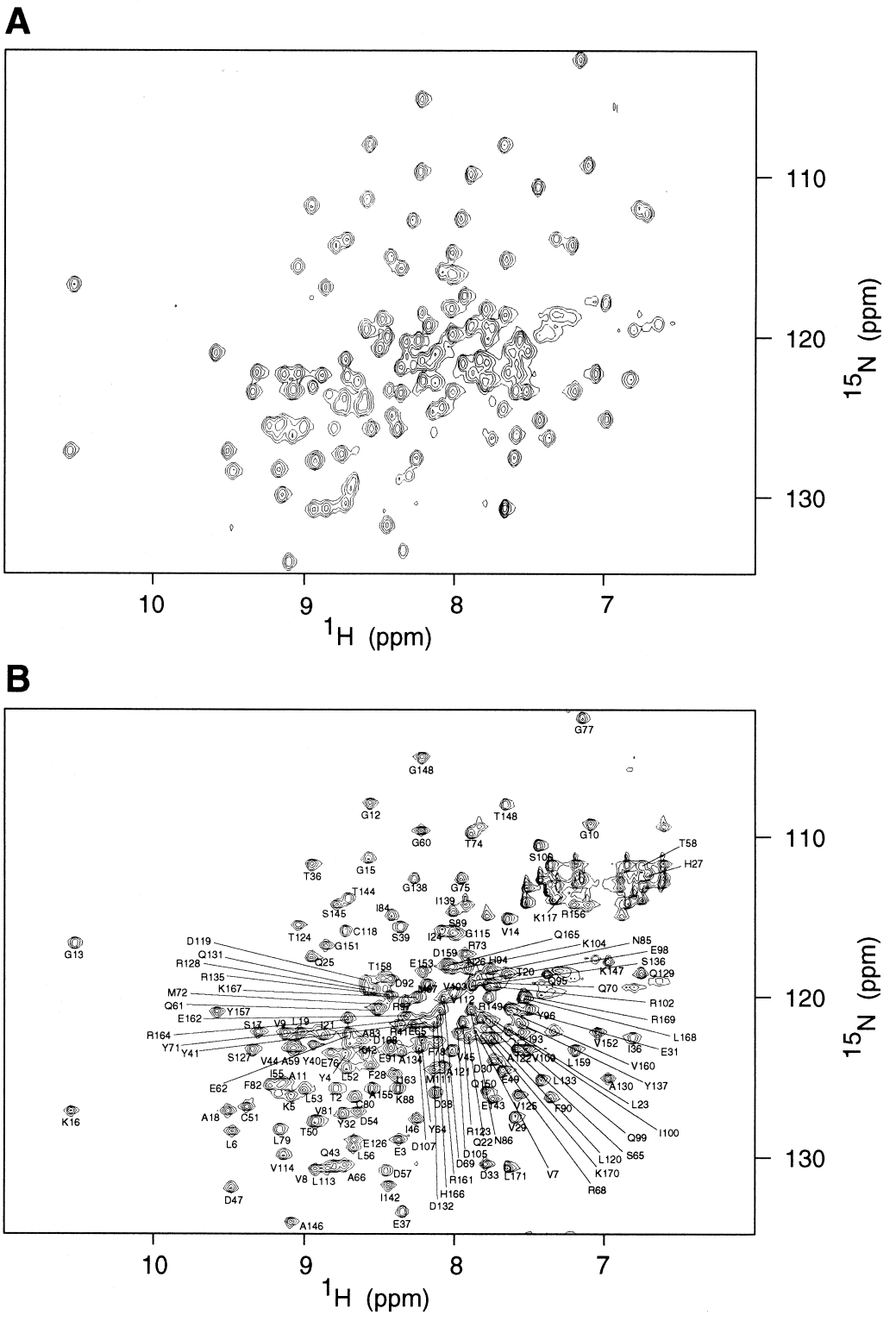


Fig. 6. The ^{15}N-HSQC spectra of the labeled Ras proteins. The backbone amide resonances and the folded sidechain amino resonances are observed. A: The Ras protein synthesized by the cell-free dialysis system with the [^{13}C,^{15}N]algal amino acid mixture. B: The Ras protein synthesized by conventional overexpression in vivo with ^{13}C-glucose and ^{15}N-ammonium chloride.

(Fig. 3). The condensed S30 extract increased the productivity of the CAT protein about 2-fold with 10 kDa and 50 kDa MWCO membrane after 6 h of incubation. The MWCO of the dialysis membrane strongly affected the production of the CAT protein. The amount of CAT protein synthesized with the 50 kDa MWCO membrane was about 3.5 times greater than that with the 10 kDa MWCO membrane, in both the original and condensed S30 extracts. With the 10 kDa MWCO membrane, the exchange of substrate seemed to be insufficient, due to membrane clogging. With the combination of the condensed S30 extract and the 50 kDa MWCO membrane, the concentration of the synthesized CAT protein

reached over 3 mg/ml in the reaction mixture. The translation kinetics of the dialysis system with the condensed S30 extract and 50 kDa MWCO membrane indicates that a high production rate was sustained for 6 h (Fig. 4A). The synthesized CAT protein was observed as a thick, major band in the SDS-PAGE analysis (Fig. 4B). The concentration of CAT protein in the external solution after 6 h of incubation was 16 μg/ml, and most of the CAT protein (MW 25 634) was retained in the internal solution, in spite of the 50 kDa MWCO. Finally, about 6 mg of the CAT protein was synthesized per ml of the reaction mixture in 21 h, by an exchange of the external solution (and supplementation of the plasmid vector and the T7 RNA polymerase) at 6 h (Fig. 5).

The amount of the CAT protein synthesized by the dialysis system with the condensed S30 extract and the 50 kDa MWCO membrane (6 mg/ml) was about 10 times larger than that with the batch system and with the 1 h incubation. The ratio of the product to the substrate in the dialysis system was comparable to that in the batch system. On the other hand, 1 μl of the original S30 extract (before condensation) produces about 1 μg of CAT protein in the batch system and about 10 μg in the dialysis system. The higher purity of the product in the reaction mixture with the dialysis system is advantageous for purification of the product protein.

The high productivity of our cell-free protein synthesis system is comparable to the productivity of the in vivo expression methods. Thus, cell-free protein synthesis will become a powerful protein production method for biochemistry and biophysics.

3.4. $^{13}C/^{15}N$-labeled Ras protein

In order to produce a $^{13}C/^{15}N$-labeled Ras protein by the cell-free system, we used $^{13}C/^{15}N$-labeled algal amino acid mixture instead of unlabeled amino acids in the reaction mixture. The algal amino acid mixture contained no cysteine, glutamine, asparagine, and tryptophan, and only small amounts of tyrosine, phenylalanine, and arginine. Therefore, these amino acids, except tryptophan, which is not contained in the Ras protein, were supplemented. Second, the potassium glutamate contained in the original reaction mixture was eliminated, because it interferes with the labeling of the glutamate residues of the Ras protein, and the other salt concentrations were optimized to eliminate the potassium glutamate. These modifications did not decrease the production of the Ras protein (data not shown).

The ^{15}N-HSQC spectrum of the labeled Ras protein synthesized in vitro was consistent with that of the uniformly labeled protein synthesized in vivo (Fig. 6). The ^{13}C-HSQC spectra were also measured successfully (data not shown). All of the observed backbone amide ^{15}N resonances for the amino acid residues, which had originally been labeled amino acids in the reaction mixture, were successfully assigned according to the resonance assignment for the uniformly labeled recombinant Ras (Fig. 6), indicating that the in vitro and in vivo preparations of the Ras protein have the same tertiary structure. As unlabeled L-glutamine, L-asparagine, and L-cysteine were used in the cell-free synthesis, the backbone amide resonances of the Asn and Cys residues and the side chain resonances of the Gln and Asn residues were missing, while the backbone amide resonances of the Gln residues were very weakly observed probably because of metabolic conversion of L-glutamate to L-glutamine (Fig. 6). In addition, the amide resonances of the Glu and Asp residues were somewhat weakened probably by metabolic dilution of stable isotopes. If all of the amino acids, including L-glutamine and L-asparagine, in the cell-free protein synthesis are labeled with stable isotopes, the Ras protein may be labeled uniformly. Therefore, cell-free protein synthesis will become a powerful protein production method for NMR spectroscopy.

Acknowledgements: We thank Drs. K. Takio and N. Dohmae (Division of Biomolecular Characterization, RIKEN, Japan) for N-terminal sequencing of Ras proteins. This work was supported in part by a grant for the Biodesign Research Program and a Special Grant for Promotion of Research, from the Institute of Physical and Chemical Research (RIKEN), Japan, Grants-in-Aid for Scientific Research from the Ministry of Education, Science, Sports and Culture, Japan (08260222, 08558076, 08780570), a Grant-in-Aid (Bio Media Program) from the Ministry of Agriculture, Forestry and Fisheries, Japan (BMP 97-V-4-2), and a Grant-in-Aid ('Research for the Future' Program) from the Japan Society for the Promotion of Science, Japan (JSPS-RFTF 96100305).

References

[1] Ellman, J., Mendel, D., Anthony-Cahill, S., Noren, C.J. and Schultz, P.G. (1991) Methods Enzymol. 202, 301–336.
[2] Sonar, S., Lee, C.P., Coleman, M., Patel, N., Liu, X., Marti, T., Khorana, H.G., RajBhandary, U.L. and Rothschild, K.J. (1994) Nature Struct. Biol. 1, 512–517.
[3] Kigawa, T., Muto, Y. and Yokoyama, S. (1995) J. Biomol. NMR 6, 129–134.
[4] Yabuki, T., Kigawa, T., Dohmae, N., Takio, K., Terada, T., Ito, Y., Laue, E.D., Cooper, J.A., Kainosho, M. and Yokoyama, S. (1998) J. Biomol. NMR 11, 295–306.
[5] Spirin, A.S., Baranov, V.I., Ryabova, L.A., Ovodov, S.Y. and Alakhov, Y.B. (1988) Science 242, 1162–1164.
[6] Kim, D.M. and Choi, C.Y. (1996) Biotechnol. Prog. 12, 645–649.
[7] Davis, J., Thompson, D. and Beckler, G.S. (1996) Promega Notes Mag. 56, 14–18.
[8] Nakano, H., Tanaka, T., Kawarasaki, Y. and Yamane, T. (1994) Biosci. Biotechnol. Biochem. 58, 631–634.
[9] Kim, D.M., Kigawa, T., Choi, C.Y. and Yokoyama, S. (1996) Eur. J. Biochem. 239, 881–886.
[10] Nakano, H., Tanaka, T., Kawarasaki, Y. and Yamane, T. (1996) J. Biotechnol. 46, 275–282.
[11] Ryabova, L.A., Vinokurov, L.M., Shekhovtsova, E.A., Alakhov, Y.B. and Spirin, A.S. (1995) Anal. Biochem. 226, 184–186.
[12] Kawarasaki, Y., Kawai, T., Nakano, H. and Yamane, T. (1995) Anal. Biochem. 226, 320–324.
[13] Ha, J.M., Ito, Y., Kawai, G., Miyazawa, T., Miura, K., Ohtsuka, E., Noguchi, S., Nishimura, S. and Yokoyama, S. (1989) Biochemistry 28, 8411–8416.
[14] Fujita-Yoshigaki, J., Ito, Y., Yamasaki, K., Muto, Y., Miyazawa, T., Nishimura, S. and Yokoyama, S. (1992) J. Protein Chem. 11, 731–739.
[15] Muto, Y., Yamasaki, K., Ito, Y., Yajima, S., Masaki, H., Uozumi, T., Walchli, M., Nishimura, S., Miyazawa, T. and Yokoyama, S. (1993) J. Biomol. NMR 3, 165–184.
[16] Ito, Y., Yamasaki, K., Iwahara, J., Terada, T., Kamiya, A., Shirouzu, M., Muto, Y., Kawai, G., Yokoyama, S., Laue, E.D., Walchli, M., Shibata, T., Nishimura, S. and Miyazawa, T. (1997) Biochemistry 36, 9109–9119.
[17] Pratt, J.M. (1984) in: (Hemes, B.D. and Higgins, S.J., Eds.), pp. 179–209, IRL Press, Oxford.
[18] Zawadzki, V. and Gross, H.J. (1991) Nucleic Acids Res. 19, 1948.
[19] Shaw, W.V. (1975) Methods Enzymol. 43, 737–755.
[20] Laemmli, U.K. (1970) Nature 227, 680–685.
[21] Odom, O.W., Kudlicki, W., Kramer, G. and Hardesty, B. (1997) Anal. Biochem. 245, 249–252.
[22] Pelham, H.R. and Jackson, R.J. (1976) Eur. J. Biochem. 67, 247–256.
[23] Choi, J., Yoon, S., Hong, H., Choi, D. and Yoo, G. (1996) Anal. Biochem. 236, 82–84.

A wheat germ cell-free system is a novel way to screen protein folding and function

EUGENE HAYATO MORITA,[1,2] TATSUYA SAWASAKI,[2,3] RIKOU TANAKA,[4] YAETA ENDO,[2,3] AND TOSHIYUKI KOHNO[4]

[1]Center for Gene Research, Ehime University, Ehime 790-8566, Japan
[2]Satellite Venture Business Laboratory, and [3]Department of Applied Chemistry, Faculty of Engineering, Ehime University, Ehime 790-8577, Japan
[4]Mitsubishi Kagaku Institute of Life Sciences (MITILS), Tokyo 194-8511, Japan

(RECEIVED November 27, 2002; FINAL REVISION March 4, 2003; ACCEPTED March 4, 2003)

Abstract

For high-throughput protein structural analysis, it is indispensable to develop a reliable protein overexpression system. Although many protein overexpression systems, such as that involving *Escherichia coli* cells, have been developed, the number of overexpressed proteins showing the same biological activities as those of the native proteins is limited. A novel wheat germ cell-free protein synthesis system was developed recently, and most of the proteins functioning in solution were synthesized as soluble forms. This suggests the applicability of this protein synthesis method to determination of the solution structures of functional proteins. To examine this possibility, we have synthesized two ^{15}N-labeled proteins and obtained ^{1}H-^{15}N HSQC spectra for them. The structural analysis of these proteins has already progressed with an *E. coli* overexpression system, and ^{1}H-^{15}N HSQC spectra for biologically active proteins have already been obtained. Comparing the spectra, we have shown that proteins synthesized with a wheat germ cell-free system have the proper protein folding and enough biological activity. This is the first experimental evidence of the applicability of the wheat germ cell-free protein synthesis system to high-throughput protein structural analysis.

Keywords: Wheat germ; cell-free; protein synthesis; HSQC; structural analysis

With the increase in the available sequence information on the genomes in various cells, attention has been turned to the structures, properties, and functional activities of proteins. However, rapid progress in the area of proteomics requires the availability of sufficient amounts of proteins. Currently, three major strategies are being used for protein production: chemical synthesis, in vivo expression, and cell-free synthesis. The first two methods have severe drawbacks. Chemical synthesis is not practical for the synthesis of long peptides (Blaschke et al. 2000), and in vivo expression can produce proteins that do not have any significant effect on the physiology of the host cells (Golf and Goldberg 1987; Chrunyk et al. 1993). With a cell-free translation system, in contrast, one can synthesize larger proteins at the same or higher speed, and as accurately as ones for in vivo translation (Kurland 1982; Pavlov and Ehrenberg 1996), and express proteins that would interfere with the host cell physiology.

One of the most convenient eukaryotic cell-free translation systems is based on wheat germ embryos containing all the components for translation in a concentrated dried state and ready for protein synthesis after germination. A past study has indicated that such systems are generally unstable and thus insufficient (Roberts and Paterson 1973). Recently, however, we found that plants contain endogeneous inhibi-

Reprint requests to: Eugene Hayato Morita, Center for Gene Research, Ehime University, 3-5-7 Tarumi, Ehime 790-8566, Japan; e-mail: ehmorita@dpc.ehime-u.ac.jp; fax: 81-89-946-9968; or Toshiyuki Kohno, Mitsubishi Kagaku Institute of Life Sciences (MITILS), 11 Minamiooyama, Machida-shi, Tokyo 194-8511, Japan; e-mail: tkohno@libra.ls.m-kagaku.co.jp; fax: 81-42-724-6296.

Abbreviation: HSQC, heteronuclear single quantum correlation spectroscopy.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.0241203.

tors of translation, and that in the case of conventional wheat germ extracts, the RNA N-glycosidase tritin and other inhibitors, such as thionin, ribonucleases, deoxyribonucleases, and proteases—those are found in the endosperm—inhibit translation (Ogasawara et al. 1999; Madin et al. 2000). Extensive washing of wheat embryos, to eliminate endosperm contamination, results in extracts with high degrees of stability and activity (Madin et al. 2000). With such an extract, the translation reaction proceeds for longer than 60 h. When performed in a dialysis bag with continuous feeding of substrates and removal of small byproducts (Spirin et al. 1988), enzymatically active proteins are yielded in milligram quantities per milliliter reaction volume (Madin et al. 2000). In the previous study, Sawasaki et al. (2002) showed the applicability of this cell-free protein synthesis system to the screening of gene products, and suggested that the expressed proteins have functions and then may attain the correct 3D structures. This applicability is crucial for modern proteomics that require a high thoroughput. However, to clarify the interrelationships between the structures and functions of proteins, it is indispensable to confirm that the functionally active proteins synthesized with this cell-free system have the proper protein folding. In this article, we report that biologically active proteins have the same solution structures as already determined. We also show that with the use of isotopically labeled amino acids as substrates, only the proteins synthesized with this cell-free system are isotopically labeled. This indicates further applicability of this cell-free system because the structural analysis of proteins can progress without any purification step when the concomitant proteins do not exhibit any strong interaction with the synthesized proteins. This property will greatly facilitate structural studies on proteins and is strongly related with the high throughput necessary for modern proteomics. The above-mentioned properties constitute the first experimental evidence that a wheat germ cell-free protein synthesis system is the best way to clarify the structure–function interrelationships of proteins.

Results

Figure 1 shows the amounts of ubiquitin and RbpA1 synthesized in 1-μL reaction mixtures determined by SDS-PAGE. The synthesized proteins are indicated by arrows. The amounts of the synthesized proteins increased significantly until 4 and 2 days incubation for ubiquitin and RbpA1, respectively. On the basis of these results, we synthesized ^{15}N labeled ubiquitin and RbpA1 for 4 and 2 days with the cell-free system, respectively. The total amounts of the synthesized proteins in 1-mL reaction mixtures were from 200 to 400 μg (from 20 to 40 nmole) and the final concentrations of NMR samples were ~40 μM. In Figures 2 and 3, the ^{1}H-^{15}N HSQC spectra of proteins overexpressed in *E. coli* cells and those synthesized with the wheat germ cell-free system are compared. On the synthesis of ubiquitin with the cell-free system, the N atoms in the side chains of Asn and Gln were not ^{15}N labeled, and the corresponding ^{15}NH signals were not observed. Almost all the backbone ^{15}NH signals overlapped. In the case of RbpA1 synthesis, on the other hand, the N atoms in the side chains of Asn and Gln were ^{15}N labeled and almost all the ^{15}NH signals overlapped. In both cases, it is indicated that the overall structures of proteins synthesized in the two different ways are almost identical.

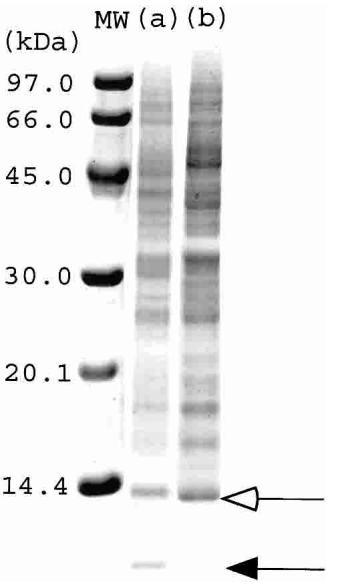


Figure 1. SDS-PAGE of reaction mixtures for (*A*) ubiquitin and (*B*) RbpA1, after synthesis. The synthesized proteins are indicated by black (for ubiquitin) and white (for RbpA1) arrows.

Furthermore, the ^{1}H-^{15}N HSQC spectra of proteins overexpressed in *E. coli* cells were for purified proteins, and those of proteins synthesized with the cell-free system for crude ones. This difference in sample conditions indicates that the concomitant proteins in the reaction mixture are not ^{15}N labeled in the process of targeted protein synthesis.

Discussion

In Figure 1, bands corresponding to the synthesized proteins can be clearly observed, and it can be estimated that the amounts of the synthesized proteins are from 200 to 400 ng/μL on the basis of the intensities of these bands. Then, the total amount of the synthesized proteins in 1-mL reaction mixture can be estimated to from 200 to 400 μg. The molecular weights of the synthesized proteins are almost 10,000, and the molar amounts of the synthesized proteins are from 20 to 40 nmole. These results are almost the same as the previous ones (Sawasaki et al. 2002).

Next, as shown in Figure 2, the ^{1}H-^{15}N HSQC spectra for the overexpressed and purified proteins are almost the same

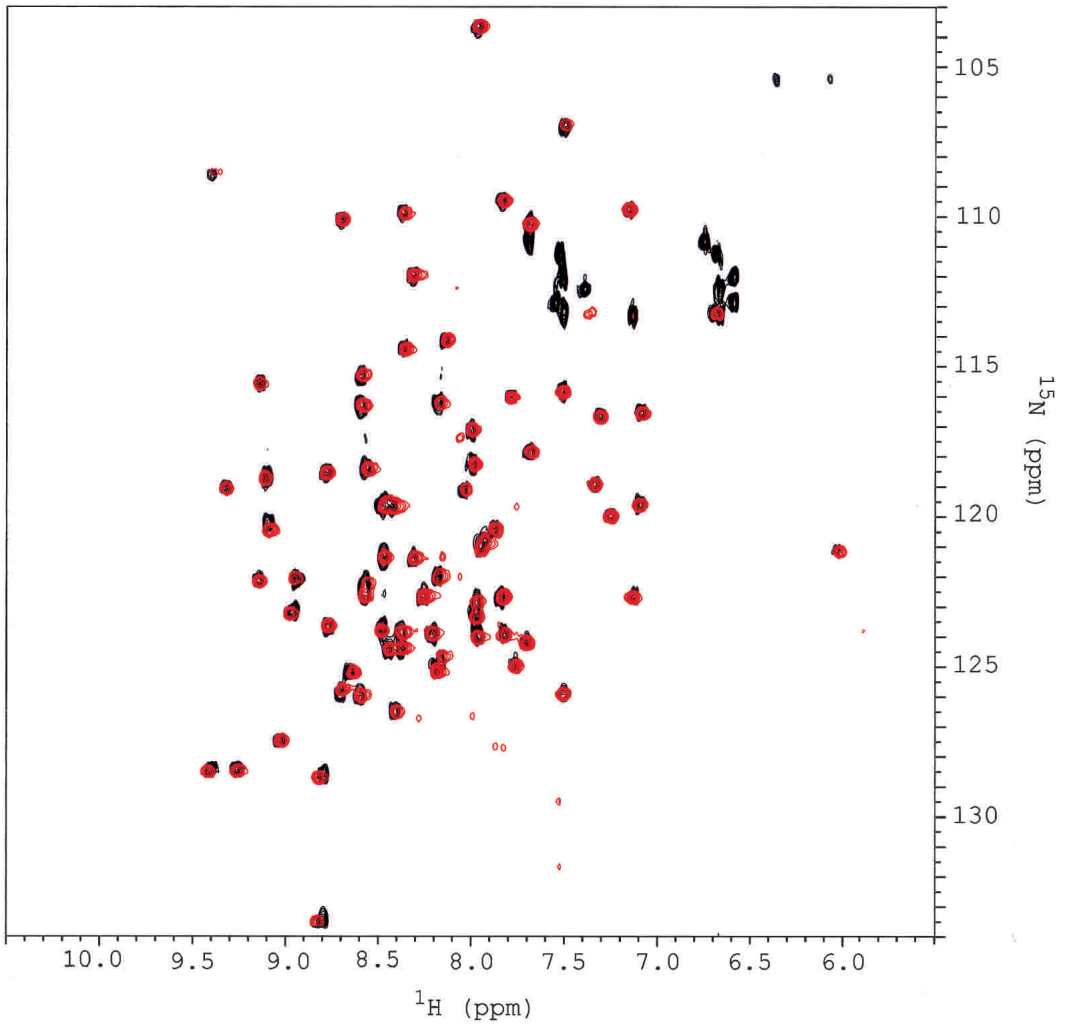


Figure 2. ^{1}H-^{15}N HSQC spectra (NMR buffer, pH 6.0, 30°C) of ^{15}N-labeled yeast ubiquitin overexpressed in *E. coli* cells (1.0 mM, 128 [t1] × 1024 [t2] complex points, 64 scans; black) and synthesized with the wheat germ cell-free system (0.10 mM, 64 [t1] × 512 [t2] complex points, 512 scans; red) were obtained at the ^{1}H resonance frequency of 500 MHz. To optimize the resolution in the nitrogen dimension, a ^{15}N spectral width of 1600 Hz was used (spectral widths of 1600 and 6250 Hz in F_1 and F_2, respectively).

as those for ones synthesized with the wheat germ cell-free system and crude proteins. This reveals an important feature of the proteins synthesized with the wheat germ cell-free system. In the ^{1}H-^{15}N HSQC spectra for the reaction mixtures, only signals corresponding to the newly synthesized proteins can be observed. This means that the amino acids added to the dialysis buffer as substrates are used to synthesize the target proteins following the genetic information of mRNA added to the reaction mixture. In the case of the cell-free protein synthesis system involving *E. coli* extracts, without purification, some signals of concomitant impurities or structural heterogeneity of synthesized proteins will be observed in the ^{1}H-^{15}N HSQC spectra of reaction mixtures (data not shown). Shimizu et al. (2001) recently reconstituted a cell-free translation system with purified components from *E. coli* extracts. However, the applicability of this system to protein structural analysis has not been shown yet. The presence of these contaminating signals may interfere with checking of the folding of synthesized proteins or monitoring of the molecular interactions between the synthesized proteins and substrates. In our case, without purification, no contaminating signal was observed in ^{1}H-^{15}N HSQC spectrum, and this feature of the wheat germ cell-free system will greatly facilitate the screening of synthesized-protein folding and determination of protein structures, compared to the case in which proteins are synthesized or overexpressed with other systems. This is one of the important features for high-throughput proteomics.

In Figure 3, the ^{1}H-^{15}N HSQC spectra of purified RbpA1 synthesized in two different ways can be seen to be quite identical. Without purification, the ^{1}H-^{15}N HSQC spectrum of RbpA1 synthesized with the wheat germ cell-free system

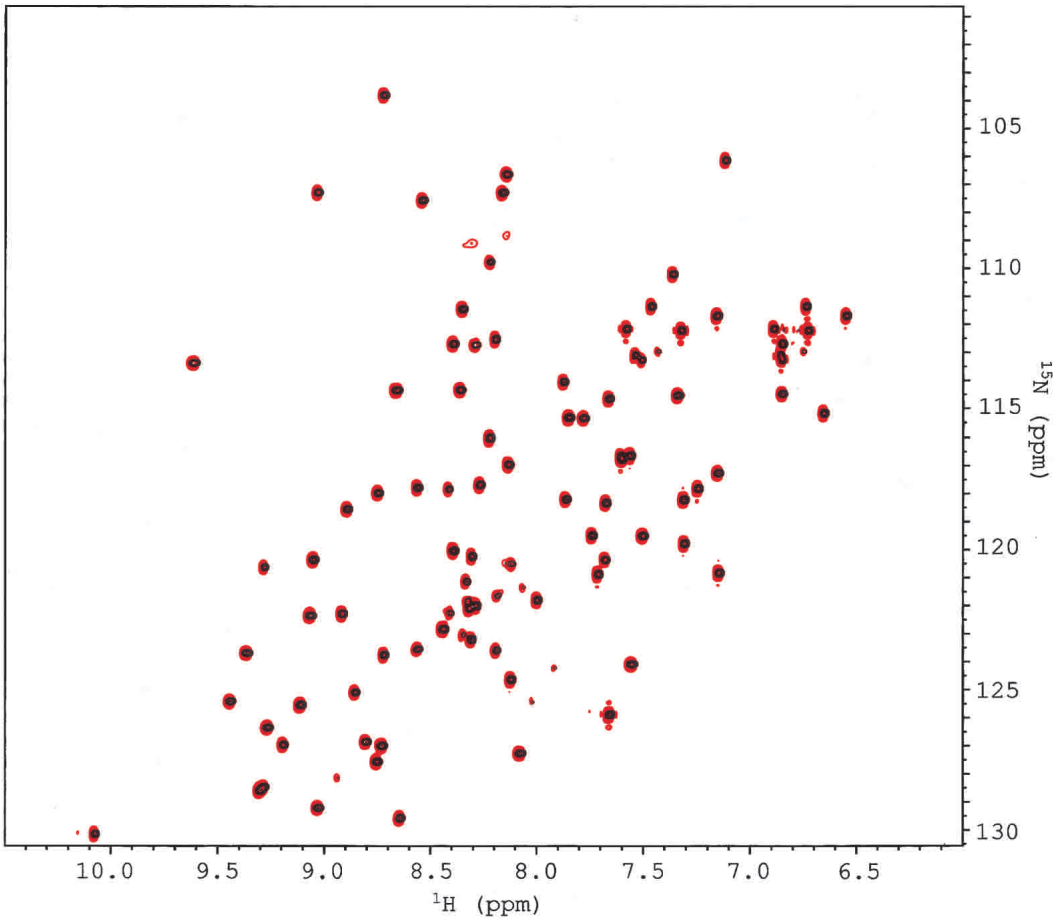


Figure 3. ^{1}H-^{15}N HSQC spectra (NMR buffer, pH 6.9, 30°C) of ^{15}N-labeled RbpA1 overexpressed in *E. coli* cells (0.5 mM, 128 [t1] × 512 [t2] complex points, 64 scans; black) and synthesized with the wheat germ cell-free system (0.12 mM, 64 [t1] × 512 [t2] complex points, 1024 scans; red) were obtained at the ^{1}H resonance frequency of 500 MHz. To optimize the resolution in the nitrogen dimension, a ^{15}N spectral width of 1500 Hz was used (spectral widths of 1500 and 8000 Hz in F_1 and F_2, respectively).

is much weaker and different from these spectra. This situation is drastically changed by the treatment of a crude RbpA1 sample with RNaseA. After such treatment, the two spectra were almost identical, which indicates that the RbpA1 interacts with the nucleic acid molecules present in the reaction mixture for the wheat germ cell-free system. However, in the case of ubiquitin, this tendency was not observed. These results revealed other important features. In the absence of concomitant molecules showing great interaction with the target proteins, the ^{1}H-^{15}N HSQC spectra of proteins overexpressed in *E. coli* and those of synthesized proteins are identical. For the proteins examined in this study, it was shown that the biological activities of these overexpressed in *E. coli* are as high as those of the native proteins. Considering these results, it can be concluded that the proteins synthesized with the wheat germ cell-free system have the proper protein folding and enough biological activity. It is further concluded that if the ^{1}H-^{15}N HSQC spectral pattern for a synthesized functionally unidentified protein changes during the purification steps, this protein must strongly interact with the concomitant molecules, which can be extracted in these steps. This means that it is easy to determine the biological function of this protein and the target molecules. From this point of view, our results can be interpreted as follows.

Ubiquitin does not interact with the concomitant molecules in a reaction mixture, and no significant spectral changes are observed between the ^{1}H-^{15}N spectra for purified and crude solutions. However, in the case of RbpA1, RbpA1 interacts with RNA molecules electrostatically, and the signal intensities observed in the ^{1}H-^{15}N HSQC spectrum are quite weak because of the increase in the apparent molecular weight. After ultracentrifugation and polyethylene imine treatment, the level of concomitant RNA molecules becomes lower and the signal intensities observed in the ^{1}H-^{15}N HSQC spectrum increase significantly. From this result, RbpA1 is thought to interact with RNA molecules strongly. As shown in a previous study (Sato 1995),

RbpA1 has inhibits the formation of double-stranded mRNA, and the spectral change observed here is in good agreement with this biological function of RbpA1.

Next, we will focus our attention to the cost of this protein synthesis. We purchased the wheat germ extract from Cell-Free Science Company. Concentration of the amino acids needed to synthesize the protein with wheat germ cell-free system is about half of that with *E. coli* cell-free system (RTS500; Roche). The total amounts obtained in 1-mL reaction mixtures with these systems are almost the same. Then, the total cost needed to synthesize fully ^{15}N-labeled proteins with wheat germ cell-free system is almost the same as that with RTS500 system. To synthesize the ^{15}N and ^{13}C double-labeled proteins, the price of each double-labeled amino acid is quite expensive, and the total cost of the *E. coli* system will be much higher than that of the wheat germ system.

In the wheat germ system, the ^{15}N-labeled amino acid mixture from the algal cell can be used as the amino acid source. In this mixture, some types of amino acids are not included, and we have to add these amino acids before use. However, the price of this mixture is quite cheap (less than a 10th of the total cost of pure ^{15}N-labeled or ^{15}N- and ^{13}C-double-labeled amino acids). Then, in the case of a uniform labeling experiment (^{15}N, ^{15}N and ^{13}C), economically, we must use the amino acid mixture from algal cells as the labeled amino acid source.

In summary, our results strongly indicate that the wheat germ cell-free protein synthesis system is one of the best ways for both protein structural analysis and the screening of the biological functions of newly found proteins.

One other important feature of the cell-free system is selective amino acid labeling of proteins. We now try to check this possibility, and despite the incompleteness of our checking, amino acids checked until now have been selectively labeled in proteins with our wheat germ system. In the near future, we will show precise information in another article.

Materials and methods

Synthesis of mRNA

The coding sequences of yeast ubiquitin and RbpA1 were amplified by the PCR method, and transferred to the EcoRV and XhoI sites of the pEU3-NII plasmid (Toyobo; T7 promoter sequence is exchanged with SP6 promoter sequence). In the presence of 16 mM Mg^{2+}, the mRNAs of these proteins were synthesized with SP6 RNA polymerase, with these plasmids as templates (Madin et al. 2000; Sawasaki et al. 2002).

Protein synthesis

Synthesized mRNAs (200 μg) were precipitated with ethanol and dissolved in 260 μL dialysis buffer, and then mixed with the wheat germ extract for protein synthesis (Madin et al. 2000). This mixture was dialyzed against the dialysis buffer containing 20 amino acids labeled with ^{15}N (Nippon Sanso) for 4 days (ubiquitin) or 2 days (RbpA1). Wheat germ extract was purchased from the Cell-Free Science Company, and other reagents (other than labeled amino acids) were purchased from Nakarai Tesque. After synthesis, the reaction mixtures were treated in two different ways, as follows.

Ubiquitin

The reaction mixture (1 mL) was concentrated to 250 μL with the use of a Centricon-3 micro-concentrator (Millipore), then the buffer was changed to NMR buffer (50 mM sodium phosphate, 100 mM NaCl, pH 6.5) by passage through a Micro Spin G-25 column (Pharmacia) equilibrated with the same buffer.

RbpA1

The reaction mixture (6 mL) was subjected to ultracentrifugation (100,000*g* for 1 h), and then polyethylene imine was added to the supernatant to the final concentration of 0.5% to remove the concomitant nucleotide derivatives. This solution was subjected to centrifugation (20,000*g* for 15 min), and then ammonium sulfate was added to the supernatant to the final concentration of 430 mg/mL. The precipitated protein was collected with a centrifuge (20,000*g* for 15 min). The obtained precipitate was suspended in 250 μL of dialysis buffer (50 mM potassium phosphate, 50 mM KCl, 1 mM EDTA, pH 6.8), and then dialyzed against the same buffer for 12 h.

Overexpression and purification of ubiquitin and RbpA1 in E. coli *cells*

Overexpression and purification of ubiquitin and RbpA1 were achieved as described previously (Sakamoto et al. 1999; Morita et al. 2000). Briefly, pET-24a and pET-21d plasmids harboring *yeast ubiquitin* and *RbpA1*, respectively, were transferred to *E. coli* BL21(DE3) cells. These cells were cultured at 37°C in M9 minimal medium containing $^{15}NH_4Cl$ and the protein expression was induced with IPTG. Cells were collected and disrupted with sonication. After centrifugation, different processes were achieved for further purification.

For yeast ubiquitin, the supernatant was incubated at 85°C for 5 min and then chilled on ice. After a further centrifugation step, the supernatant was further purified with ion-exchange column chromatography (SP Sepharose FF column; Pharmacia) and gel-filtration chromatography (HiLoad 26/60 Superdex 75 pg column; Pharmacia).

For RbpA1, the supernatant was purified with ion-exchange and gel-filtration column chromatography only (Q Sepharose FF column, and HiLoad 26/60 Superdex 75 pg column; Pharmacia).

The purified proteins were concentrated and applied to the NMR study.

Measurement of HSQC spectra

D_20 was added to the concentrated protein solutions (final concentration, 10%). HSQC spectra of these proteins were obtained with a DMX 500 (Bruker) FT-NMR spectrometer. The data were processed using NMRPipe (Delaglio et al. 1995) on a Linux work-

station. The ^{1}H and ^{15}N chemical shifts were referenced according to the method of Wishart et al. (1995).

Acknowledgments

This work was supported by a grant from National Project on Protein Structural and Functional Analyses.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Blaschke, U.K., Silberstein, J., and Muir, T.W. 2000. Protein engineering by expressed protein ligation. *Methods Enzymol.* **328:** 478–496.

Chrunyk, B.A., Evans, J., Lillquist, J., Young, P., and Wetzel, R. 1993. Inclusion body formation and protein stability in sequence variants of interleukin-1 β. *J. Biol. Chem.* **268:** 18053–18061.

Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J., and Bax, A. 1995. NMR Pipe: A multdimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6:** 277–293.

Golf, S.A. and Goldberg, A.L. 1987. An increased content of protease La, the lon gene product, increases protein degradation and blocks growth in *Escherichia coli*. *J. Biol. Chem.* **262:** 4508–4515.

Kurland, C.G. 1982. Translational accuracy in vitro. *Cell* **28:** 201–202.

Madin, K., Sawasaki, T., Ogasawara, T., and Endo, Y. 2000. A highly efficient and robust cell-free protein synthesis system prepared from wheat embryos: Plants apparently contain a suicide system directed at ribosomes. *Proc. Natl. Acad. Sci.* **97:** 559–564.

Morita, E.H., Murakami, T., Uegaki, K., Yamazaki, T., Sato, N., Kyogoku, Y., and Hayashi, H. 2000. NMR backbone assignments of the cold-regulated RNA-binding protein, RbpA1, in the cyanobacterium, *Anabaena variabilis* M3. *J. Biomol. NMR* **17:** 351–352.

Ogasawara, T., Sawasaki, T., Morishita, R., Ozawa, A., Madin, K., and Endo, Y. 1999. A new class of enzyme acting on damaged ribosomes: Ribosomal RNA apurinic site specific lyase found in wheat germ. *EMBO J.* **18:** 6522–6531.

Pavlov, M.Y. and Ehrenberg, M. 1996. Rate of translation of natural mRNAs in an optimized in vitro system. *Arch. Biochem. Biophys.* **328:** 9–16.

Roberts, B.E. and Paterson, B.M. 1973. Efficient translation of tobacco mosaic virus RNA and rabbit globin 9S RNA in a cell-free system from commercial wheat germ. *Proc. Natl. Acad. Sci.* **70:** 2330–2334.

Sakamoto, T., Tanaka, T., Ito, Y., Rajesh, S., Iwamoto-Sugai, M., Kodera, Y., Tsuchida, N., Shibata, T., and Kohno, T. 1999. An NMR analysis of ubiquitin recognition by yeast ubiquitin hydrolase: Evidence for novel substrate recognition by a cysteine protease. *Biochemistry* **38:** 11634–11642.

Sato, N. 1995. A family of cold-regulated RNA-binding protein genes in the cyanobacterium *Anabaena variabilis* M3. *Nucleic Acid Res.* **23:** 2161–2167.

Sawasaki, T., Ogasawara, T., Morishita, R., and Endo, Y. 2002. A cell-free protein synthesis system for high throughput proteomics. *Proc. Natl. Acad. Sci.* **99:** 14652–14657.

Shimizu, Y., Inoue, A., Tomari, Y., Suzuki, T., Yokogawa, T., Nishikawa, K., and Ueda, T. 2001. Cell-free translation reconstituted with purified components. *Nat. Biotechnol.* **19:** 751–755.

Spirin, A.S., Baranov, V.I., Ryabova, L.A., Ovodov, S.Y., and Alakhov, Y.B. 1988. A continuous cell-free translation system capable of producing polypeptides in high yield. *Science* **25:** 1162–1164.

Wishart, D.S., Bigam, C.G., Yao, J., Abildgard, F., Dyson, H.J., Oldfield, E., Markley, J.L., and Sykes, B.D. 1995. ^{1}H, ^{13}C and ^{15}N chemical shift referencing in biomolecular NMR. *J. Biomol. NMR* **6:** 135–140.

Isotope filtered/edited NMR methods for the study of biomolecular complexes - Acquisition & Optimization

Ruediger Weisemann

Bruker Biospin GmbH, Silberstreifen, D-76275 Rheinstetten, Germany
ruediger.weisemann@bruker-biospin.de

Since the study of biomolecular complexes is becoming a more and more important topic for molecular biology, and as NMR is the most powerful tool to do these studies, the motivation of this lecture is to provide an introduction into the terminology used to discriminate between different classes of experiments, into their basic principles and finally into "how things are done" in real life, i.e. on the spectrometer.

The common denominator of quite a few of these pulse sequences is the use of adiabatic pulses for the inversion of ^{13}C magnetization under specific conditions. Starting with a recapitulation of the basic properties of adiabatic pulses and their usage in modern pulse sequences, the design of the pulses used in isotope-filtered experiments will be discussed and their properties will be analysed based on simulations. From these simulations, parameter settings can be derived for the pulse sequences in the Bruker standard experiment library.

In order to evaluate typical experimental results on a "real-life" sample, examples of several experiments have been acquired using a Cryoprobe^{TM} on a 500 MHz instrument and will be discussed in connection with the pulse sequences employed.

Related Literature:

1) A. Breeze, *Prog. NMR Spectrosc.* **36** (2000), 323-372
2) Zwahlen et al., *J.Am.Chem.Soc*. **119** (1997), 6711-6721
3) Andersen, P. and Otting, G., *J.Magn.Reson*. **144** (2000), 168-170
4) Gruschus, J.M. and Ferreti, J.A., *J.Magn.Reson*. **135** (1998), 87-92
5) Iwara, J. et al., *J.Biomol.NMR* **19** (2001), 231-241
6) Lee, W. et al., *FEBS Letters* **350** (1994), 87-90
7) Ogura et al., *J. Biomol.NMR* **8** (1996), 492-498
8) Pascal, S.M. et al., *J.Magn.Reson* **B 103** (1994), 197-201
9) Uhrín, D. et al., *J. Biomol.NMR* **18** (2000), 253-259

Multidimensional NMR spectral processing and experimental scheme for NMR structural genomics.

Youlin Xia[1], Xiaolian Gao[1] and Guang Zhu[1,2]

[1]Department of Chemistry, University of Houston, Houston, TX 77004-5003
[2]Dept of Biochemistry, The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong, SAR, China

GFT-NMR can be applied to rapidly obtain NMR spectral assignments of istopically labeled proteins. Here, ten (3, 2)D GFT-NMR experiments were demonstrated with $^{13}C/^{15}N$-labeled ubiquitin sample on a Bruker Avance 800 spectrometer. This set of experiments can be done in about 18 hours, while the necessary experiments for backbone assignments, (3, 2)D HNCACB, HN(CO)CACB, HNCO, and 2D ^{15}N-^{1}H HSQC can be done in less than 6 hours. With (3, 2)D GFT-NMR experiments, the experimental times are considerably reduced compared with that of conventional 3D experiments. (3, 2)D GFT-NMR experiments can be easily performed and analyzed. It is believed that this set of experiments will speed up the structural determination of proteins by NMR. In addition, linear prediction methods can also be extensively used for extrapolating FIDs to obtain higher resolution spectra with less experimental times.

November 1st

Handout for laboratory

(B1A Conference Room)

NMR Workshop Lab Session

Group No. 組別	Name 姓名	Institute 單位	Registration No. 報名序號
1	游春愛	IBMS, Academia Sinica	004
1	方珮如	IBMS, Academia Sinica	021
1	徐駿森	IBC, Academia Sinica	055
2	Sue, Shih-Che	IBMS, Academia Sinica	022
2	Chingyu Chou	IBMS, Academia Sinica	035
2	Yi-Jan Lin	IBMS, Academia Sinica	069
3	Iren Wang	IBMS, Academia Sinica	001
3	羅元超	IBMS, Academia Sinica	006
3	陳金榜	IBMS, Academia Sinica	025
4	林達顯	National Yang-Ming U.	002
4	Yi-Choang Huang	National Yang-Ming U.	059
4	Chi-jen Lo	National Yang-Ming U.	060
5	Ko-Hsin Chin	National Chunghsing U.	017
5	Shan-Ho Chou	National Chunghsing U.	024
5	蔡文評	IBC, Academia Sinica	041
6	Yu-nan Liu	National Tsing Hua U.	014
6	ChengChao-sheng	National Tsing Hua U.	030
6	Shangwu Ding	National Sun Yat-sen U.	054
7	魏明財	Chem, Academia Sinica	005
7	吳丞偉	Chem, Academia Sinica	015
7	Shu-Chuan Jao	Chem, Academia Sinica	066
8	Chia-Lin Chyan	National Dong Hwa U.	038
8	C. G. Sudhahar	National Chunghsing U.	046
8	Aranganathan	National Tsing Hua U.	056
9	Wen-Yih Jeng	National Cheng Kung U.	003
9	Woei-Jer Chuang	National Cheng Kung U.	032
9	Luo Shih Chi	IBC, Academia Sinica	039
10	Tjong Siu Cin	National Tsing Hua U.	068
10	魏淑怡	National Tsing Hua U.	071
10	Ya-Ping Tsao	National Tsing Hua U.	074

November 2nd

Handout for lecture

(B1C Auditorium)

Journal of Biomolecular NMR, **17:** 125–136, 2000.
KLUWER/ESCOM

A tracked approach for automated NMR assignments in proteins (TATAPRO)

H.S. Atreya, S.C. Sahu, K.V.R. Chary* & Girjesh Govil
Department of Chemical Sciences, Tata Institute of Fundamental Research, Homi Bhabha Road, Colaba, Mumbai 400005, India

Received 25 January 2000; Accepted 18 April 2000

Key words: automated NMR assignments, *Borrelia burgdoferi* OspA, drosophila numb phosphotyrosine-binding domain, *Eh*-CaBP, *Escherichia coli* maltose binding protein, fibroblast collagenase, sequence specific resonance assignments, triple resonance experiments

Abstract

A novel automated approach for the sequence specific NMR assignments of $^{1}H^{N}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'/^{1}H^{\alpha}$ and ^{15}N spins in proteins, using triple resonance experimental data, is presented. The algorithm, TATAPRO (**T**racked **A**u**T**omated **A**ssignments in **Pro**teins) utilizes the protein primary sequence and peak lists from a set of triple resonance spectra which correlate $^{1}H^{N}$ and ^{15}N chemical shifts with those of $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'/^{1}H^{\alpha}$. The information derived from such correlations is used to create a '*master_list*' consisting of all possible sets of $^{1}H^{N}_{i}$, $^{15}N_{i}$, $^{13}C^{\alpha}_{i}$, $^{13}C^{\beta}_{i}$, $^{13}C'_{i}/^{1}H^{\alpha}_{i}$, $^{13}C^{\alpha}_{i-1}$, $^{13}C^{\beta}_{i-1}$ and $^{13}C'_{i-1}/\,^{1}H^{\alpha}_{i-1}$ chemical shifts. On the basis of an extensive statistical analysis of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shift data of proteins derived from the BioMagResBank (BMRB), it is shown that the 20 amino acid residues can be grouped into eight distinct categories, each of which is assigned a unique two-digit code. Such a code is used to tag individual sets of chemical shifts in the master_list and also to translate the protein primary sequence into an array called pps_array. The program then uses the master_list to search for neighbouring partners of a given amino acid residue along the polypeptide chain and sequentially assigns a maximum possible stretch of residues on either side. While doing so, each assigned residue is tracked in an array called assig_array, with the two-digit code assigned earlier. The assig_array is then mapped onto the pps_array for sequence specific resonance assignment. The program has been tested using experimental data on a calcium binding protein from *Entamoeba histolytica* (*Eh*-CaBP, 15 kDa) having substantial internal sequence homology and using published data on four other proteins in the molecular weight range of 18–42 kDa. In all the cases, nearly complete sequence specific resonance assignments ($> 95\%$) are obtained. Furthermore, the reliability of the program has been tested by deleting sets of chemical shifts randomly from the master_list created for the test proteins.

Introduction

Sequence specific resonance assignments (hereafter abbreviated as ssr_assignments) in proteins are an important and essential step towards complete three dimensional (3D) structural characterization (Wüthrich et al., 1986). In recent years, several double and triple resonance experiments have been proposed to carry out ssr_assignments in isotope labeled proteins (Bax and Grzesiek, 1993). However, for large proteins, manual assignment becomes a tedious and time consuming task. This has led to an increasing demand of the development of algorithms for automation of ssr_assignments, following which a number of strategies have been proposed (see review by Moseley and Montelione, 1999). These include approaches which utilize information from various triple resonance experiments and methods such as simulated annealing

*To whom correspondence should be addressed. E-mail: chary@tifr.res.in

(Buchler et al., 1997; Lukin et al., 1997), bayesian statistics and artificial intelligence (Zimmerman et al., 1997; Montelione et al., 1999), characteristic $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts of individual amino acid residues (Grzesiek and Bax, 1993; Friedrichs et al., 1994), threshold accepting algorithm (Leutner et al., 1998), connectivity tracing algorithms (Olson and Markley, 1994) and neural networks (Choy et al., 1993; Hare and Prestegard, 1994). For side chain assignments, methods have been proposed which utilize side chain topologies of spin systems (Li and Sanctuary, 1997), side chain ^{13}C chemical shift patterns of amino acid residues (Zimmerman et al., 1994) and semi-automated approaches (Meadows et al., 1994). Other strategies utilize information from homologous proteins and chemical shift prediction for complete ssr_assignments (Bartels et al., 1996; Gronwald et al., 1999).

In this paper, we propose a novel algorithm for automated ssr_assignments of $^{1}H^{N}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'/^{1}H^{\alpha}$ and ^{15}N spins in proteins, called TATAPRO, using the protein primary sequence and a set of triple resonance experiments which correlate the $^{1}H^{N}$ and ^{15}N chemical shifts with those of $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'/^{1}H^{\alpha}$. This approach is demonstrated using data from CBCANH (Wittekind and Mueller, 1993), CBCA(CO)NH (Grzesiek and Bax, 1992), HN(CA)CO (Clubb et al., 1992a) and HNCO (Kay et al., 1990) spectra.

Two important parameters which determine the success of any automated approach for ssr_assignments in proteins are good input of peaks, both in terms of resolution and sensitivity, and a reliable classification of individual spin systems (Moseley and Montelione, 1999). With the development of a TROSY-based approach for the implementation of various triple resonance pulse sequences (Salzmann et al., 1998, 1999) and also with the modification of pulse sequences for deuterated proteins (Gardner and Kay, 1998), it is now possible to acquire triple resonance spectra with high resolution and sensitivity for proteins with molecular weights up to 50 kDa (Loria et al., 1999). These techniques, combined with an efficient peak picking algorithm, help in overcoming ambiguities in resonance assignments to some extent. However, a satisfactory classification of individual spin systems is still an important and difficult task. Few of the algorithms in the past have utilized the characteristic $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts of individual amino acid residues (Grzesiek and Bax, 1993; Friedrichs et al., 1994; Lukin et al., 1997) for such a classification. However, due to extensive overlap of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts for most of the amino acid residues, identifying each residue by its characteristic chemical shifts can result in ambiguous assignments. The problem is further aggravated by unusual chemical shifts which can lead to erroneous assignments. Hence, in order to obtain insight into the distribution of NMR chemical shifts for amino acid residues, we have carried out an extensive statistical analysis using $^{13}C^{\alpha}$ ($\sim$25 000) and $^{13}C^{\beta}$ ($\sim$21 000) chemical shift information of all the proteins deposited in the BMRB (Seavey et al., 1991). This analysis aided in grouping the 20 amino acid residues into eight distinct categories, based on their characteristic $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts. These categories are then distinguished from each other by assigning them a unique two-digit code. This grouping of amino acid residues dramatically reduces the problem of overlapping and unusual chemical shifts and results in a deterministic approach to the problem of ssr_assignments.

The algorithm has been tested for assignments, using experimental data, on a calcium binding protein from *Entamoeba histolytica* (*Eh*-CaBP, M_r $\sim$15 kDa), which possesses a substantial internal sequence homology, and on four other proteins with published assignments. The complete ssr_assignments have been accomplished in three stages, using a separate program at each stage. These programs have been written in ANSI C code and can be compiled on any Unix-based workstation or Windows-based system equipped with a C compiler. The execution time of the program is of the order of a few seconds on an R10000-based solid impact workstation (SGI). The program can be obtained on request at the following e-mail address: chary@tifr.res.in

Methodology

$^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shift statistics

The algorithm for ssr_assignments proposed here primarily makes use of the characteristic $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts of each individual amino acid residue except Pro residues. For this purpose, an extensive statistical analysis has been carried out, utilizing the chemical shift data available in the BMRB.

The histograms in Figures 1a and 1b depict the percentage of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts, respectively, spanning the range 12–78 ppm (28–78 ppm for $^{13}C^{\alpha}$) for individual amino acid residues. As evident from Figure 1a, Gly($^{13}C^{\alpha}$) always resonates upfield of 50 ppm in a region well separated from the $^{13}C^{\alpha}$ chemi-

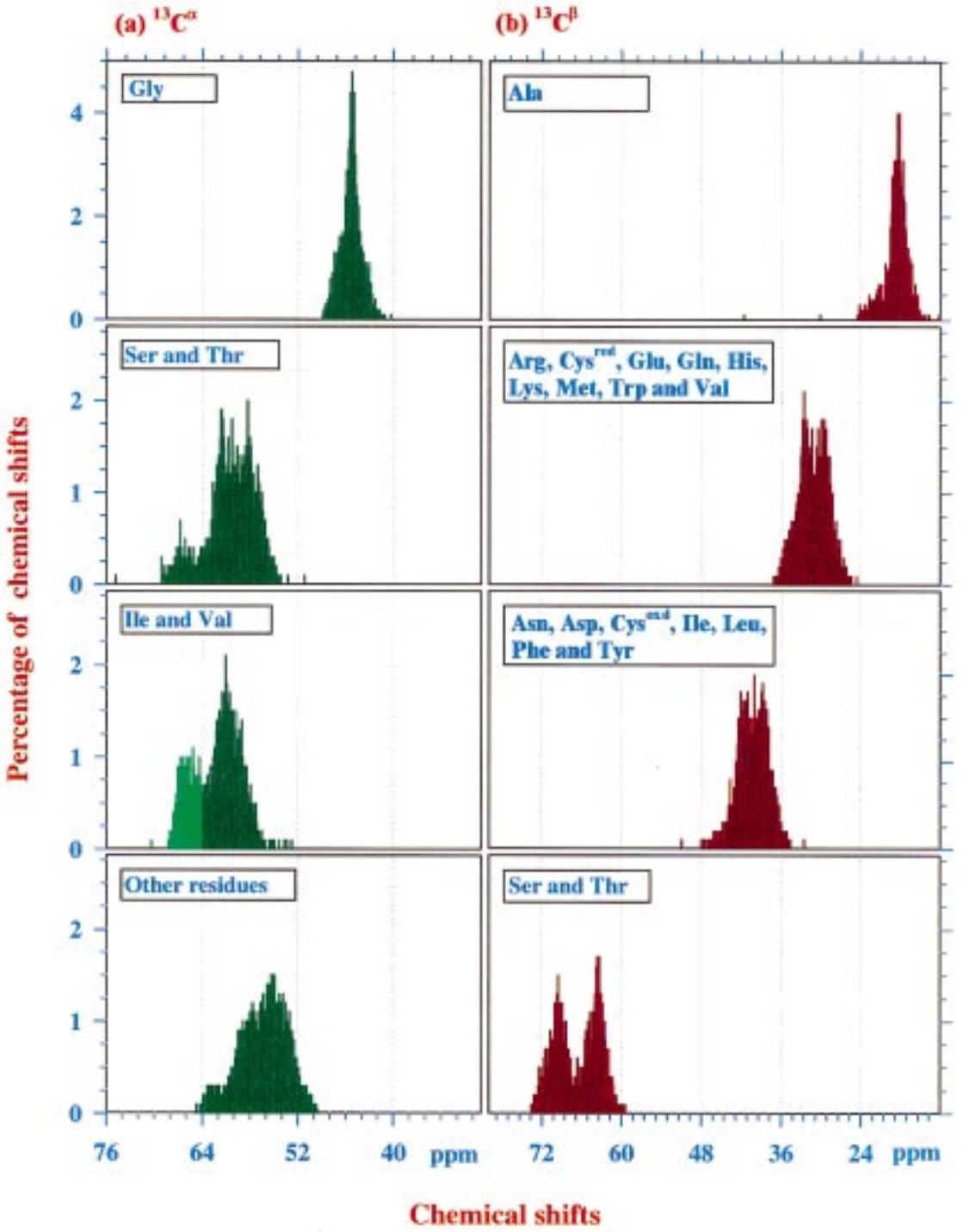


Figure 1. Distribution of (a) $^{13}C^{\alpha}$ and (b) $^{13}C^{\beta}$ chemical shifts for various amino acid residues in proteins selected from the BMRB. The histograms depict the percentage of amino acids having a particular chemical shift within a range of 0.1 ppm. In the case of Ile and Val residues, $^{13}C^{\alpha}$ chemical shifts greater than 64 ppm are shown in a different colour for clarity.

cal shifts of all other residues. On the other hand, no $^{13}C^{\beta}$ resonates between 50–58 ppm (Figure 1b). It is also seen that the amino acid residues can be classified into five distinct categories based entirely on the characteristic $^{13}C^{\beta}$ chemical shifts (Figure 1b): (i) Gly having no $^{13}C^{\beta}$; (ii) less than 24 ppm – Ala; (iii) 24–36 ppm – Arg, Cys^{red}, Gln, Glu, His, Lys, Met, Val, Trp; (iv) 36–50 ppm – Asp, Asn, Cys^{oxd}, Ile, Leu, Phe and Tyr; and (v) more than 58 ppm – Ser and Thr.

In our algorithm, each of these five categories is distinguished from the rest by a single digit code. For example, all Gly residues are given a code **1**, Ala residues **2** and so on (Table 1). Besides Ser and Thr residues, Val and Ile residues (about 26% of them) also have their $^{13}C^{\alpha}$ chemical shifts downfield of 64 ppm (Figure 1a). No other amino acid residue has resonances in this region. This facilitates a further classification for Val and Ile residues by appending a second digit to the single digit codes assigned earlier. This second digit is chosen as **1** for all the residues with $^{13}C^{\alpha}$ chemical shifts downfield of 64 ppm and $^{13}C^{\beta}$ chemical shift upfield of 58 ppm and is chosen as **0** otherwise. Thus, for example, a Val residue with its $^{13}C^{\alpha}$ chemical shift downfield of 64 ppm acquires a code **4 1**, while a Val residue with its $^{13}C^{\alpha}$ chemical shift upfield of 64 ppm acquires a code **4 0** (Table 1).

The last two columns in Table 1 indicate the percentage of amino acid residues which violate the two-digit code assigned to them. This might happen if a given residue exhibits unusual $^{13}C^{\alpha}$ or/and $^{13}C^{\beta}$ chemical shift(s), as a result of which it acquires a code different from the one generally expected. For example, it is evident from Table 1 that Ser, Thr and Ala residues deviate the least from their expected range of $^{13}C^{\beta}$ chemical shifts and rarely do other amino acid residues fall in their range. Further, since these three residues, along with Gly, are given individual codes, it is easy to identify these spin systems uniquely. Hence, these four residues serve as primary markers in ssr_assignments, as has been observed earlier (Metzler et al., 1993).

Experimental inputs for TATAPRO

Several automated assignment strategies that have been proposed in the past require correlating peak lists from a large number (six or more) of triple resonance experiments (Friedrich et al., 1994; Olson and Markley, 1994; Zimmerman et al., 1995; Lukin et al., 1997). These strategies suffer from the fact that there are likely to be some chemical shift variations for the same spin in different spectra because of changes in experimental conditions such as pH and decoupling heating during the experiments. Also, 4D experiments required as input for some of these algorithms suffer from low digital resolution and sensitivity. These factors can contribute to incomplete or/and erroneous assignments. Further, for proteins with low stability, it is imperative that all data be acquired in a short duration of time. Thus, it is desirable to restrict the number of experiments required as inputs to a minimum and all experiments should be performed under identical conditions of pH and temperature, preferably with the same sample. In the present study, four triple resonance experiments, namely CBCANH, CBCA(CO)NH, HNCO and HN(CA)CO, are found to be sufficient for complete ssr_assignment of all the $^{1}H^{N}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$ and ^{15}N spins. Other 3D triple resonance experiments which can serve as inputs for TATAPRO are HN(CA)HA (Clubb et al., 1992b) and HN(COCA)HA (Clubb and Wagner, 1992) in place of HN(CA)CO and HNCO, respectively. Peak lists obtained from these spectra consisting of chemical shift co-ordinates of the peaks, $(\omega_1, \omega_2, \omega_3) = (^{13}C/^{1}H^{\alpha}, ^{15}N, ^{1}H^{N})$, along with their intensities and phases, are taken as inputs for TATAPRO.

Description of the algorithm

We have considered a deterministic approach here which takes into account the characteristic $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts of all the 20 individual amino acid residues. The approach can be divided into three important steps, namely, peak list preparation, assignment of two-digit codes (Table 1) to the individual amino acid residues in the primary sequence and the rows in the *master_list* and finally, carrying out ssr_assignments. These steps are described below:

(a) Peak list preparation. Peak lists derived from CBCANH, CBCA(CO)NH, HNCO and HN(CA)CO (or alternatively, HN(CA)HA and HN(COCA)HA spectra) are used to group the chemical shifts as follows. An automatically picked CBCA(CO)NH peak list has information about $^{13}C^{\alpha}_{i-1}$ and $^{13}C^{\beta}_{i-1}$ chemical shifts for a given pair of $^{15}N_i$ and $^{1}H^{N}_{i}$. From such a list, the chemical shifts of $^{13}C^{\alpha}_{i-1}$ and $^{13}C^{\beta}_{i-1}$ are identified for each specific pair of $^{15}N_i$ and $^{1}H^{N}_{i}$ chemical shifts within the user defined tolerance limits and grouped into a single set. Owing to the fact that all peaks in CBCA(CO)NH are seen with positive intensity, distinction between those arising from $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ is based on the following criteria:

Table 1. Two-digit codes assigned to different amino acid residues based on their characteristic $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shift ranges. The last two columns indicate the percentage of residues which violate these codes

Sr. no.	$^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts (δ in ppm) characteristics	Amino acids	Two-digit code	Percentage of $^{13}C^{\beta}$ chemical shift violations	Percentage of other residues taking the code
1	Absence of $^{13}C^{\beta}$	Gly	**1 0**	0.0	0.00
2	$15 < \delta(^{13}C^{\beta}) < 24$	Ala	**2 0**	0.8	0.09
3	$\delta(^{13}C^{\beta}) > 58$	Ser and Thr	**3 0**	0.5	0.04
4	$24 < \delta(^{13}C^{\beta}) < 36$ & $\delta(^{13}C^{\alpha}) < 64$	Lys, Arg, Gln, Glu, His, Trp, Cys^{red}, Val and Met	**4 0**	3.1	2.4
5	$24 < \delta(^{13}C^{\beta}) < 36$ & $\delta(^{13}C^{\alpha}) \geq 64$	Val	**4 1**	1.3	0.6
6	$36 < \delta(^{13}C^{\beta}) < 50$ & $\delta(^{13}C^{\alpha}) < 64$	Asp, Asn, Phe, Tyr, Cys^{Oxd}, Ile and Leu	**5 0**	3.0	2.4
7	$36 < \delta(^{13}C^{\beta}) < 50$ & $\delta(^{13}C^{\alpha}) \geq 64$	Ile	**5 1**	6.8	0.3
8	–	Pro	**6 0**	–	–

(i) If the ^{13}C chemical shift of one of the peaks at ($^{13}C_{i-1}$, $^{15}N_i$, $^{1}H_i^N$) is below 50 ppm and the other is more than 50 ppm, then the former is treated as due to $^{13}C^{\beta}$ and the latter to $^{13}C^{\alpha}$.
(ii) If the ^{13}C chemical shifts of both the peaks ($^{13}C_{i-1}$, $^{15}N_i$, $^{1}H_i^N$) are more than 50 ppm, then they belong to Ser/Thr residues. Since either peak may be due to the $^{13}C^{\alpha}$ or $^{13}C^{\beta}$ spin, two possible combinations of chemical shifts are considered.
(iii) If only one peak ($^{13}C_{i-1}$, $^{15}N_i$, $^{1}H_i^N$) is seen with its ^{13}C chemical shift below 50 ppm, then it is categorically treated as due to Gly($^{13}C^{\alpha}$) and the corresponding $^{13}C^{\beta}$ chemical shift is set to zero.
(iv) In the event of degeneracy in $^{15}N_i$ and $^{1}H_i^N$ chemical shifts for $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ peaks, all possible combinations of chemical shifts are chosen.

On the other hand, an automatically picked CBCANH peak list has information about $^{13}C_i^{\alpha}$, $^{13}C_i^{\beta}$, $^{13}C_{i-1}^{\alpha}$ and $^{13}C_{i-1}^{\beta}$ chemical shifts for a given pair of $^{15}N_i$ and $^{1}H_i^N$ chemical shifts. For each set of $^{15}N_i$, $^{1}H_i^N$, $^{13}C_{i-1}^{\alpha}$ and $^{13}C_{i-1}^{\beta}$ chemical shifts grouped from the CBCA(CO)NH spectral peak list, the search is now carried out in the CBCANH peak list to identify $^{13}C_i^{\alpha}$ and $^{13}C_i^{\beta}$ chemical shifts, within the user defined tolerance limits. In principle, for a given pair of $^{15}N_i$ and $^{1}H_i^N$ chemical shifts, one observes four (^{13}C, $^{15}N_i$, $^{1}H_i^N$) peaks (for non-Gly residues): one pair belonging to ($^{13}C_i^{\alpha}$/$^{13}C_i^{\beta}$, $^{15}N_i$, $^{1}H_i^N$) peaks and the other to ($^{13}C_{i-1}^{\alpha}$/$^{13}C_{i-1}^{\beta}$, $^{15}N_i$, $^{1}H_i^N$) peaks. Hence, it may seem straightforward to identify ($^{13}C_i^{\alpha}$, $^{15}N_i$, $^{1}H_i^N$) and ($^{13}C_i^{\beta}$, $^{15}N_i$, $^{1}H_i^N$) peaks which are positive and negative in intensity respectively, given that the sequential peaks ($^{13}C_{i-1}^{\alpha}$/$^{13}C_{i-1}^{\beta}$, $^{15}N_i$, $^{1}H_i^N$) are already identified. In practice, due to overlap in cross peaks ($^{13}C^{\alpha}$/$^{13}C^{\beta}$,$^{15}N_i$, $^{1}H_i^N$) of self and sequential residues, one may not find four distinct peaks. Our algorithm then makes use of the following criteria for identifying $^{13}C_i^{\alpha}$ and $^{13}C_i^{\beta}$ peaks:
(i) Since only Gly($^{13}C^{\alpha}$) spins resonate below 50 ppm, all peaks (^{13}C, ^{15}N, $^{1}H^N$) with positive intensity and ^{13}C chemical shift below 50 ppm are ignored (Figure 1a).
(ii) Since no $^{13}C^{\beta}$ spin resonates between 50–58 ppm, all peaks (^{13}C, ^{15}N, $^{1}H^N$) with negative intensity and ^{13}C chemical shift within this range are ignored (Figure 1b).
(iii) The most intense positive peak, excluding the one belonging to the sequential residue, is identified as the ($^{13}C_i^{\alpha}$, $^{15}N_i$, $^{1}H_i^N$) peak.
(iv) The most intense negative peak, excluding the one belonging to the sequential residue, is identified as the ($^{13}C_i^{\beta}$, $^{15}N_i$, $^{1}H_i^N$) peak.
(v) If the absolute intensity of either or both of the ($^{13}C_i^{\alpha}$/$^{13}C_i^{\beta}$, $^{15}N_i$, $^{1}H_i^N$) peaks happens to be less than twice the intensity of the corresponding sequential peak, or if no peak other than the sequential peaks ($^{13}C_{i-1}^{\alpha}$/$^{13}C_{i-1}^{\beta}$, $^{15}N_i$, $^{1}H_i^N$) is seen, then the sequential peaks are themselves treated as those for the self ($^{13}C_i^{\alpha}$/$^{13}C_i^{\beta}$, $^{15}N_i$, $^{1}H_i^N$) peaks. This can happen if either or both of the ($^{13}C^{\alpha}$/$^{13}C^{\beta}$, ^{15}N, $^{1}H^N$) peaks of self and sequential residues are degenerate.

Once $^{1}H_i^N$, $^{15}N_i$, $^{13}C_i^{\alpha}$, $^{13}C_i^{\beta}$, $^{13}C_{i-1}^{\alpha}$ and $^{13}C_{i-1}^{\beta}$ chemical shifts are grouped into individual sets, $^{13}C_i'$ and $^{13}C_{i-1}'$ chemical shifts are obtained using automatically picked HN(CA)CO and HNCO peak lists, respectively. Thus, such grouping of chemical shifts results in a peak list containing individual sets of $^{1}H_i^N$, $^{15}N_i$, $^{13}C_i^{\alpha}$, $^{13}C_i^{\beta}$, $^{13}C_i'$, $^{13}C_{i-1}^{\alpha}$, $^{13}C_{i-1}^{\beta}$, and $^{13}C_{i-1}'$ chemical shifts. This list, referred to as the '*master_list*', forms the input for the next step in our algorithm.

Each individual set of chemical shifts in the master_list will hereafter be referred to as a *row*. In principle, the number of rows should correspond to the number of amino acid residues in the protein minus the number of Pro residues. In practice, owing to the near degeneracy in $^{1}H^N$ and ^{15}N chemical shifts, all possible pairs of chemical shifts within the user defined tolerance limits are accounted for in the master_list. Hence the number of rows usually exceeds the number of amino acid residues. In the case of *Eh*-CaBP with 134 residues, the master_list contained 216 rows.

(b) Assignment of two-digit codes. As discussed earlier, we classify the amino acid residues into eight different categories based on their characteristic $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts, rather than characterizing them individually as has been done in the past. This method of classification helps in a deterministic approach for resonance assignment. The two-digit code assigned to individual amino acid residues (Table 1) is used to tag the individual rows in the master_list depending on the observed $^{13}C_i^{\alpha}$ and $^{13}C_i^{\beta}$ chemical shift values. In the next step, the master_list is rearranged such that rows belonging to Gly residues are grouped together at the beginning of the list, followed by Ala etc., in the same order as in Table 1. When a polypeptide stretch of amino acid residues is assigned, the two-digit codes associated with the individual rows in that stretch are put into an array, referred to as *assig_array.* Simultaneously, all the amino acid residues in the protein primary sequence are assigned the two-digit code given in Table 1 (Ile and Val residues are given codes **4 1** and **5 1**, respectively). All Cys residues in the primary sequence are assigned a code **5 0**, corresponding to the oxidized state. The reduced Cys residues in the protein then, can be considered as having unusual chemical shifts, which are still assigned unambiguously. However, if most of the Cys residues present in the protein under investigation are in the reduced form, the user can interactively assign these residues a code **4 0**. Thus, on assigning these codes to all the individual amino acid residues in the protein primary sequence, it gets translated into an array of two-digit codes referred to as *pps_array*.

(c) Sequence specific resonance assignment. The algorithm uses the master_list for ssr_assignments. As described earlier, each row in the master_list consists of $^{1}H_i^N$, $^{15}N_i$, $^{13}C_i^{\alpha}$, $^{13}C_i^{\beta}$, $^{13}C_i'$, $^{13}C_{i-1}^{\alpha}$, $^{13}C_{i-1}^{\beta}$, and $^{13}C_{i-1}'$ chemical shift values. To begin with, the algorithm reads in the $^{13}C_i^{\alpha}$, $^{13}C_i^{\beta}$ and $^{13}C_i'$ chemical shift values from the first row in the master_list and searches for a row where, within the user-defined tolerance limits, these three chemical shifts are seen as $^{13}C_{i-1}^{\alpha}$, $^{13}C_{i-1}^{\beta}$, and $^{13}C_{i-1}'$ chemical shifts. If the search is successful, the two-digit code associated with the new row is stored in an assig_array. This procedure corresponds to forward assignment in the primary sequence, which is continued until a break is encountered. The break can be due to a Pro residue, (a) missing peak(s) or the fact that the C-terminal end of the polypeptide chain has been reached. Once a stretch of amino acid residues has been assigned in the forward direction, the algorithm continues with the assignment in the backward direction starting again from the first row in the master_list. For backward assignment, the program reads in the $^{13}C_{i-1}^{\alpha}$, $^{13}C_{i-1}^{\beta}$, and $^{13}C_{i-1}'$ chemical shifts for a given row in the master_list and searches for the row where these chemical shifts are seen as $^{13}C_i^{\alpha}$, $^{13}C_i^{\beta}$ and $^{13}C_i'$ chemical shifts, within the user-defined tolerance limits. If the search is successful, the two-digit code associated with the new row is stored in the same assig_array, as was done in the case of forward assignment. The assignment is continued until a break is encountered. Thus, after assigning the residues in both forward and backward directions, the program maps the assig_array onto the *pps_array*. A one-to-one correspondence with the pps_array results in the sequence specific resonance assignment of that polypeptide stretch. Following this, all the assigned rows are deleted from the master_list before the next round of assignment commences, for which the first row in the updated master_list is chosen as the next starting point. In principle, the above procedure suffices to assign all the amino acid residues in the protein except the Pro residues. In practice, however, several problems can arise when assigning and mapping a stretch of amino acid residues onto the pps_array. We consider each of these in detail:

(i) During the assignment procedure, more than one possible pair of $^{1}H_{i}^{N}$ *and* $^{15}N_{i}$ *chemical shifts satisfy the assignment condition.* The program continues with the assignment along each possible pathway until a break is encountered. Each of the assigned polypeptide stretches, represented in the form of a specific assig_array, is then mapped onto the pps_array. If only one of the assig_arrays gets mapped uniquely onto the pps_array, the rest of the assig_arrays are ignored and resonance assignment is continued. If more than one assig_array correspond to different stretches in the pps_array, no assignment is carried out and all the rows are retained in the master_list. The algorithm then continues the assignment with the next top row in the master_list as the starting point.

(ii) An assigned stretch of amino acid residues occurs more than once in the primary sequence. This happens mostly if the assigned polypeptide stretch of amino acid residues is of a short length (2–4 residues). In the case of proteins with substantial internal sequence homology, larger stretches (5–6 residues) are also found to be redundant. An insight in the statistics of short stretch *polypeptide redundancies* (2–8 residues in length) has been obtained by scanning eight proteins of different lengths ranging from 134 to 370 amino acid residues (Supplementary material, Table 1). In all these proteins, the number of polypeptide redundancies (2–7 residues in length) increases when the 20 amino acid residues are grouped into eight distinct categories compared to when they are considered independently. However, the amino acid stretches comprising eight or more residues are found to be unique. Such stretches can thus be assigned unambiguously. If mapping of assig_array onto pps_array results in multiple matches, the polypeptide stretch is not considered to be assigned and the assignment is continued with the next upper-most row in the master_list. Once a large fraction of amino acid residues are assigned, the number of polypeptide redundancies reduces considerably, leading to unambiguous assignment of stretches spanning even two to three residues.

(iii) One or more residues in the assigned stretch have unusual $^{13}C^{\alpha}$ *or* $^{13}C^{\beta}$ *chemical shifts and therefore do not belong to their expected category.* For example, a Val residue with $^{13}C^{\beta} = 24$ ppm may acquire a code corresponding to an Ala residue. In such a situation, referred to as a '*mismatch*', the mapping of assig_array onto the pps_array will result in either an incorrect mapping or no mapping. For polypeptide stretches spanning eight residues or more, incorrect mapping is unlikely, as these stretches will be unique in the primary sequence (Supplementary material, Table 1). In view of this, the program assigns polypeptide stretches of eight or more residues without a limit on the number of mismatches. For polypeptide stretches of 4–7 residues in length, the program allows only two mismatches, while for stretches spanning 2–3 residues, only one mismatch is allowed. At all stages of assignments, mismatches are reported to the user along with their ^{13}C chemical shift(s).

A statistical survey of $^{13}C^{\beta}$ chemical shifts in the 100 proteins chosen from the BMRB (accession numbers of proteins are listed in the Supplementary material, Table 2) reveals that in a given protein, on average the maximum number of mismatches is 2.3. Figure 2a shows the number of mismatches observed in each of the 100 proteins. Further, three proteins with the largest number of mismatches were analyzed to check the positions of these mismatches along the primary sequence. As shown in Figure 2b, the mismatches in a protein are generally distributed throughout the primary sequence and it is unlikely for a given polypeptide stretch of less than 10 residues to have more than two mismatches. This implies that such a stretch with two mismatches can still be mapped uniquely onto the primary sequence.

(iv) Assignment of a lone residue flanked by two polypeptide segments. During the process of ssr_assignments described above, one may end up with several unassigned lone residues other than prolines, that are flanked by assigned polypeptide stretches. This can happen either because of degenerate chemical shifts or due to the absence of a ($^{13}C_{i}$, $^{15}N_{i}$, $^{1}H_{i}^{N}$) peak in the respective triple resonance spectra. In such an event, the information about the $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ chemical shifts of the residue preceding the unassigned one is used to assign the ^{15}N and ^{1}H chemical shifts of the latter by utilizing CBCA(CO)NH and HNCO peak lists. Thus, by following this procedure, ^{15}N and ^{1}H chemical shifts of all the lone residues except prolines are assigned unambiguously. On the other hand, the $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ chemical shifts of all unassigned lone residues and those of Pro residues are obtained from the row corresponding to their succeeding residue in the master_list.

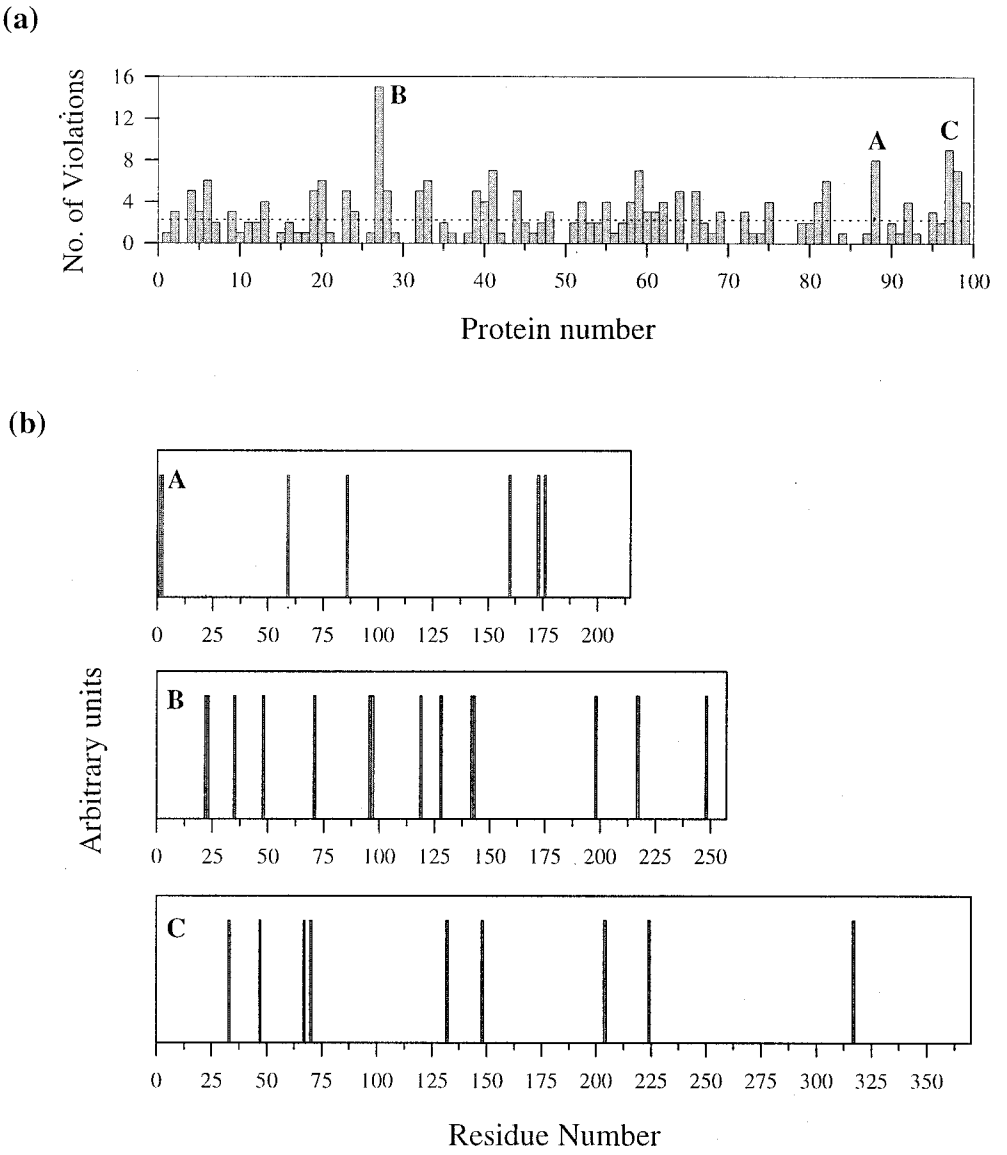


Figure 2. (a) Number of mismatches (amino acid residues which violate our classification of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ and chemical shifts, see text) observed in each of the 100 proteins chosen for statistical analysis. The dotted line indicates the average number of mismatches. (b) Mismatch locations in the primary sequences of three proteins, marked as A, B and C in (a). The respective BMRB accession numbers are: A – 4318, B – 4076 and C – 4354.

Results and discussion

Assignments in Eh-CaBP using experimental data

The algorithm has been tested for NMR assignments in *Eh*-CaBP (134 amino acid residues, 15 kDa) using the experimental data. Ssr_assignments for this protein have been reported elsewhere (Sahu et al., 1999), and were utilized to check the results using TATAPRO. *Eh*-CaBP has the characteristic EF-hands of calcium binding proteins possessing substantial internal sequence homology within the four calcium binding loops. This is evident from its primary sequence shown below, where highly homologous loop segments are highlighted:

MAEALFKEI**DVNGDGAVSYEE**VKAFVSKKRAIKNEQLLQ
LIFKSI**DADGNGEIDQNE**FAKFYGSIQGQDLSDDKIGLK
VLYKLM**DVDGDGKLTKEE**VTSFFKKHGIEKVAEQVMKA
DANGDGYITLEEFLEFSL

Table 2. Details of test proteins and percentage of assignments obtained in each case

Sr. no.	Proteins	BMRB accn. no.	Mol. wt (in kDa)	No. of amino acid residues	No. of mismatches[a]	No. of Pro residues	Percentage of assignments obtained on random deletion of peaks		
							0%	15%	30%
1	Calcium binding protein from *Entamoeba histolytica*	4271	15	134	3	0	96	85	77
2	Drosophila numb phosphotyrosine binding domain	4263	17.8	160	11	6	100	89	68
2	Fibroblast collagenase	4064	18.7	169	0	11	100	87	70
3	*Borrelia burgdorferi* OspA	4076	28	257	14	1	100	92	74
4	*Escherichia coli* maltose binding protein	4354	42	370	9	21	100	90	65

[a]Number of mismatches include reduced cysteines, as all cysteines are given a code **5 0** corresponding to the oxidized form (see text).

Such internal sequence homology complicates the resonance assignment. First, it results in multiple matches when an assigned stretch of amino acid residues in the loop region is mapped onto the primary sequence. Secondly, chemical shifts of ($^{13}C_i^{\alpha}$/$^{13}C_i^{\beta}$/$^{13}C_i'$) spins belonging to similar residues in the different loop regions are generally degenerate. This results in more than one pathway for the assignment along the polypeptide chain. Both situations can result in erroneous assignments. However, TATAPRO helps in overcoming these problems, as discussed below.

Automatically picked peak lists were obtained using the software Felix97 (Molecular Simulations Inc., San Diego, CA) from the four 3D triple resonance spectra, CBCANH, CBCA(CO)NH, HNCO and HN(CA)CO. Peaks were picked with a low threshold in all the spectra to avoid missing real peaks with low intensity, particularly in CBCANH and HN(CA)CO spectra. However, peaks from CBCA(CO)NH and HNCO experimental spectra were picked at a higher threshold, because of their inherent higher sensitivity. Thus, for 134 residues, about 3500 peaks were picked in the CBCANH spectrum and 1144 peaks in the CBCA(CO)NH spectrum. The corresponding figures for HNCO and HN(CA)CO spectra were 172 and 299, respectively.

Starting with a tolerance limit of 0.01 ppm along the $^{1}H^{N}$ dimension and 0.05 ppm along the ^{15}N dimension, chemical shifts obtained from these spectra were grouped and re-arranged to form a master_list containing rows of $^{1}H_i^{N}$, $^{15}N_i$, $^{13}C_i^{\alpha}$, $^{13}C_i^{\beta}$, $^{13}C_i'$, $^{13}C_{i-1}^{\alpha}$, $^{13}C_{i-1}^{\beta}$ and $^{13}C_{i-1}'$ chemical shifts. Wherever the ($^{13}C_i$, $^{15}N_i$, $^{1}H_i^{N}$) peaks were not found within this tolerance limit, the tolerance was gradually increased until a set of cross peaks for $^{13}C^{\alpha}$ or/and $^{13}C^{\beta}$ was seen. Next, by beginning at the first row in the master_list, which belonged to a Gly residue ($\delta(^{13}C_i^{\beta}) = 0.0$), sequence specific resonance assignment was carried out using 0.5 ppm as the tolerance limit for $^{13}C^{\alpha}$ chemical shifts, 0.2 ppm for $^{13}C^{\beta}$ chemical shifts and 0.025 ppm for $^{13}C'$ chemical shift. Once the percentage of assigned residues reached around 75%, these tolerance limits were automatically increased to 1.0 ppm, 0.4 ppm and 0.05 ppm, respectively. Both these tolerance limits and the number of residues to be assigned in a single run can be interactively altered by the user, depending on the requirement. Following this procedure, about 95% of the residues could be assigned sequence specifically. The exceptions were M1, A2, E3, I9, N56, Y62, G76 and E111. In the final stage of the algorithm, residues N56, Y62, G76 and E111, each of which were flanked by two assigned polypeptide stretches, were assigned unambiguously. Cross peaks ($^{13}C_i$, $^{15}N_i$, $^{1}H_i^{N}$) for M1, A2 and E3 were not observed in any of the aforementioned triple resonance spectra, while cross peaks for I9 could not be picked in the CBCANH spectrum at the chosen threshold. In the present case, there were three mismatches corresponding to V17 ($\delta(^{13}C_i^{\beta}) = 22.5$ ppm), L57 ($\delta(^{13}C_i^{\beta}) = 32.60$ ppm) and K94 ($\delta(^{13}C_i^{\beta}) = 39.7$ ppm). Polypeptide stretches containing these mismatches could still be assigned uniquely. Figure 3 shows the order in which different polypeptide stretches of amino acid residues were assigned during subsequent runs.

To test the robustness of the program, rows in the master_list were deleted randomly (up to 30%)

Figure 3. Polypeptide stretches assigned during subsequent runs for *Eh*-CaBP. The numbers in the boxes indicate the order in which these polypeptide stretches have been assigned using TATAPRO. The starting residue within a stretch is indicated by a black dot above it. The hashed boxes indicate lone residues flanked by two assigned polypeptide stretches.

and assignments were carried out, as discussed above. However, the tolerance limits for ^{13}C chemical shifts were increased after about 50% of the residues got assigned. The percentages of assignment obtained after the deletion of 15% and 30% of the rows were 85% and 77%, respectively (Table 2). When a large number of rows are deleted, only short stretches of amino acid residues are assigned (3–4 residues). This results in multiple mapping of assig_array onto the pps_array for such a stretch, leading to a decline in the percentage of unambiguous assignments. Further, we have observed that keeping a high ^{13}C chemical shift tolerance limit (> 0.5 ppm) in the beginning, results in erroneous assignments, as it leads to incorrect assignment pathways. This is more likely when a large number of rows of chemical shifts are deleted from the master_list. Thus, to start with, it is recommended to keep the ^{13}Cchemical shift tolerance as low as possible (about 0.25 ppm for $^{13}C^{\alpha}$, 0.15 ppm for $^{13}C^{\beta}$ and 0.1 ppm for $^{13}C'$). The tolerance limits can be increased automatically when a large number of residues ($\sim$75%) are assigned, as the chances of assignment proceeding along an incorrect pathway then reduce considerably.

Assignments using published data

BMRB data for four other proteins was used to test the reliability of our approach. These proteins, namely, drosophila numb phosphotyrosine-binding domain complexed with a phosphotyrosine peptide (17.8 kDa), a fragment of fibroblast collagenase (18.7 kDa), *Borrelia burgdoferi* OspA (28 kDa) and *Escherichia coli* maltose binding protein (42 kDa), either have a large number of mismatches (Table 2) or have a high degree of polypeptide redundancies within their primary sequence (Supplementary material, Table 1). Further, two of these proteins, namely, fibroblast collagenase and *Escherichia coli* maltose binding protein, possess a large number of Pro residues, restricting the assignable polypeptide stretches to shorter fragments, which in turn can lead to multiple mapping. Thus, resonance assignments in these four proteins constituted a rigorous evaluation for the reliability of our algorithm.

Ssr_assignments in the test proteins were carried out as in the case of *Eh*-CaBP. However, since the data sets were perfect, relatively narrow tolerance limits corresponding to 0.3 ppm for $^{13}C^{\alpha}$ chemical shifts, 0.2 ppm for $^{13}C^{\beta}$ chemical shifts and 0.025 ppm for $^{13}C'$ chemical shifts were chosen for the test proteins. First, all the rows in the master_list created for each protein were retained. This resulted in 100% ssr_assignment of the residues for which assignments have been reported. In the case of drosophila numb phosphotyrosine-binding domain, there are nine Cys

residues with all their $\delta(^{13}C_i^{\beta}) < 36$ ppm, which reflects the reduced state of these residues. In view of this, all the Cys residues were initially given a code **4 0**, which resulted in 100% unambiguous resonance assignments. In order to evaluate the robustness of our approch, the process was repeated with all the Cys residues assigned a code **5 0**, as mentioned earlier. It is interesting that this did not affect the percentage of assignments.

To further verify the reliability of the program, several rows from the master_list were randomly deleted (up to 30%) and the assignments were repeated. The percentage of assignments obtained in each case is shown in Table 2. With 15% random deletion of rows from the master_list, about 90% resonance assignment is achieved in the test cases. For proteins having a large number of Pro residues (which constitute a break during assignments in our algorithm), the percentage of assignments obtained after 30% deletion of peaks declined to 65%. On the other hand, it is observed that a large number of polypeptide segmental redundancies within the primary sequence or a large number of mismatches do not affect the percentage of assignments obtained, establishing the reliability of this approach.

Conclusions

The approach adopted here for resonance assignments resembles to some extent the one proposed by Friedrichs et al. (1994). However, there are subtle methodological differences and improvements. (i) Our algorithm is based on a deterministic approach (as opposed to probabilistic). (ii) Only four triple resonance experiments are required as input to our algorithm. (iii) We classify the 20 amino acid residues into eight different categories, with each category having a characteristic $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shift range. This approach, which has been found to be more useful and reliable, is facilitated by a two-digit code. (iv) In the event of an unexpected degeneracy in all the $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ chemical shifts, our approach explores along all possible pathways to maximize the stretch of amino acid residues that can be assigned. (v) No manual intervention is required to check the grouping of peaks or to check the residues which do not satisfy their characteristic $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shift range. The latter is checked and reported to the user automatically. The assignments of test proteins in the molecular weight range of 15–42 kDa, wherein assignments up to 75% are achieved even after 30% random deletion of peaks, establish the reliability of the program. Further, the program is shown to be robust to internal sequence homology and unusual chemical shifts. Thus, TATAPRO can be succesfully used for the assignment of large-size proteins.

Acknowledgements

The facilities provided by the National Facility for High Field NMR, supported by the Department of Science and Technology (DST), the Department of Biotechnology (DBT), the Council of Scientific and Industrial Research (CSIR), and the Tata Institute of Fundamental Research, Mumbai, are gratefully acknowledged.

References

Bartels, C., Billeter, M., Güntert, P. and Wüthrich, K. (1996) *J. Biomol. NMR*, **7**, 207–213.

Bax, A. and Grzesiek, S. (1993) *Acc. Chem. Res.*, **26**, 131–138.

Buchler, N.E.G., Zuiderweg, E.R.P., Wang, H. and Goldstein, R.A. (1997) *J. Magn. Reson.*, **125**, 34–42.

Choy, W.Y., Sanctuary, B.C. and Zhu, G. (1997) *J. Chem. Inf. Comput. Sci.*, **37**, 1086–1094.

Clubb, R.T., Thanabal, V. and Wagner, G. (1992a) *J. Magn. Reson.*, **97**, 213–217.

Clubb, R.T., Thanabal, V. and Wagner, G. (1992b) *J. Biomol. NMR*, **2**, 203–210.

Clubb, R.T. and Wagner, G. (1992) *J. Biomol. NMR*, **2**, 389–394.

Friedrichs, M.S., Mueller, L. and Wittekind, M. (1994) *J. Biomol. NMR*, **4**, 703–726.

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

Gardner, K.H., Zhang, X., Gehring, K. and Kay, L.E. (1998) *J. Am. Chem. Soc.*, **120**, 11738–11748.

Gronwald, W., Willard, L., Jellard, T., Boyko, R.F., Rajarathnam, K., Wishart, D.S., Sönnichsen, F.D. and Sykes, B.D. (1998) *J. Biomol. NMR*, **12**, 395–405.

Grzesiek, S. and Bax, A. (1992) *J. Am. Chem. Soc.*, **114**, 6291–6293.

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

Hare, B.J. and Prestegard, H. (1994) *J. Biomol. NMR*, **4**, 35–46.

Ikura, M., Kay, L.E. and Bax, A. (1990) *Biochemistry*, **29**, 4659–4667.

Kay, L.E., Ikura, M., Tschudin, R. and Bax, A. (1990) *J. Magn. Reson.*, **89**, 496–514.

Li, K.-B. and Sanctuary, B.C. (1997) *J. Chem. Inf. Comput. Sci.*, **37**, 467–477.

Li, S.C., Zwahlen, C., Vincent, S.J., McGlade, C.J., Kay, L.E., Pawson, T. and Forman-Kay, J.D. (1998) *Nat. Struct. Biol.*, **5**, 1075–1083.

Leutner, M., Gschwind, R.M., Liermann, J., Schwarz, C., Gemmecker, G. and Kessler, H. (1998) *J. Biomol. NMR*, **11**, 31–43.

Loria, J.P., Rance, M. and Palmer, A.G. (1999) *J. Magn. Reson.*, **141**, 180–184.

Lukin, J.A., Gove, A.P., Talukdar, S.N. and Ho, C. (1997) *J. Biomol. NMR*, **9**, 151–166.

Meadows, R.P., Olejniczak, E.T. and Fesik, S.W. (1994) *J. Biomol. NMR*, **4**, 79–96.
Metzler, W.J., Constantine, K.L., Friedrichs, M.S., Bell, A.J., Ernst, E.G., Lavoie, T.B. and Mueller, L. (1993) *Biochemistry*, **32**, 13818–13829.
Montelione, G.T., Rios, C.B., Swapna, G.V.T. and Zimmerman, D.E. (1999) In *Biological Magnetic Resonance, Volume 17: Structure, Computation and Dynamics in Protein NMR* (Eds., Krishna, R. and Berliner, L.), Plenum Press, New York, NY, pp. 81–130.
Moseley, H.N.B. and Montelione, G.T. (1999) *Curr. Opin. Struct. Biol.*, **9**, 635–642.
Moy, F.J., Pisano, M.R., Chandra, P.K., Urbano, C., Killar, L.M., Sung, M.L. and Powers, R. (1997) *J. Biomol. NMR*, **10**, 9–19.
Olson Jr., J.B. and Markley, J.L. (1994) *J. Biomol. NMR*, **4**, 385–410.
Pham, T.-N. and Koide, S. (1998) *J. Biomol. NMR*, **11**, 407–414.
Sahu, S.C., Atreya, H.S., Chauhan, S., Bhattacharya, A., Chary, K.V.R. and Govil, G. (1999) *J. Biomol. NMR*, **14**, 93–94.
Salzmann, M., Pervushin, K., Wider, G., Senn, H. and Wüthrich, K. (1998) *Proc. Natl. Acad. Sci. USA*, **95**, 13585–13590.
Salzmann, M., Wider, G., Pervushin, K., Senn, H. and Wüthrich, K. (1999) *J. Am. Chem. Soc.*, **121**, 844–848.
Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J.L. (1991) *J. Biomol. NMR*, **1**, 217–236.
Wittekind, M. and Mueller, L. (1993) *J. Magn. Reson.*, **B101**, 201–205.
Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York, NY.
Zimmerman, D.E., Kulikowski, C., Wang, L.L., Lyons, B.A. and Montelione, G.T. (1994) *J. Biomol. NMR*, **4**, 241–256.
Zimmerman, D.E., Kulikowski, C.A., Huang, Y., Feng, W., Tashiro, M., Shimotakahara, S., Chien, C., Powers, R. and Montelione, G.T. (1997) *J. Mol. Biol.*, **269**, 592–610.

values has indicated that for BPR and CatV ligands, chelation of $[Co(1,10\text{-phen})_2]^{2+}$ moiety occurs from 3′, 4′ position of the quinone ring (see Scheme 4). Hence, on the TiO_2 surface anchoring process takes place through the gallol and catechol rings of the BPR and CatV ligands respectively. The surface anchoring observed for the two complexes is quite similar to the anchoring which occurs in the case of anthrocyanine dye material[2].

The findings clearly suggest that non-planarity of the triphenylmethane-type ligand strongly influences the observed photocurrent conversion efficiencies. This is reflected by the different intensities of the observed MLCT transitions for the complexes. Variation in the strength of the MCLT transition occurs as a result of different torsion angles. In the case of high torsion angles substantial *dp–p** interactions take place, and similarly, in the case of low torsion angles less *dp–p** interactions occur.

1. Desilvestro, J., Gratzel, M., Kavan, L. and Moser, J., *J. Am. Chem. Soc*., 1985, **107**, 2988–2992.
2. (a) Tennakone, K., Kumara, G. R. R. A., Kottegoda, I. R. M., Wijayantha, K. G. U. and Perera, V. P. S., *J. Phys. D*: *Appl. Phys*., 1998, **31**, 1492–1495; (b) Nazeeruddin, M. K. *et al*., *J. Am. Chem. Soc*., 1993, **115**, 6382–6386.
3. Cherepy, N. J., Smestad, G. P., Gratzel, M. and Zhang, J. Z., *J. Phys. Chem. B*, 1997, **101**, 9342–9345.
4. Jayaweera, P. M., Kumarasinghe, A. R. and Tennakone, K., *J. Photochem. Photobiol. A*, 1999, **126**, 111–114.
5. Jayaweera, P. M., Palayangoda, S. S. and Tennakone, K., *J. Photochem. Photobiol. A*, 2001, **140**, 173–176.
6. Burns, D. T. and Dadger, D., *Analyst*, 1980, **105**, 1082–1086.
7. Vinodgopal, K., Hua, X., Dahlgren, R. L., Lappin, A. G., Patterson, L. K. and Kamat, P. V., *J. Phys. Chem*., 1995, **99**, 10883–10887.
8. Zerner, M. C., *Inorg. Chem*., 1986, **28**, 2728–2733.

ACKNOWLEDGEMENT. We are grateful to the National Science Foundation for support.

Received 20 March 2002; revised accepted 25 September 2002

Automated NMR assignments of proteins for high throughput structure determination: TATAPRO II

H. S. Atreya, K. V. R. Chary* and Girjesh Govil

Department of Chemical Sciences, Tata Institute of Fundamental Research, Homi Bhabha Road, Colaba, Mumbai 400 005, India

TATAPRO (Tracked Automated Assignments in Proteins), a novel algorithm for automated NMR assignments in proteins is presented to aid in high throughput protein structure determination. In this version (TATAPRO II), the 20 amino acid residues are classified into nine distinct categories based on their characteristic $^{13}C^{a}$ and $^{13}C^{b}$ chemical shifts derived statistically using a database of ~ 100,000 shifts. Further, in the current version, the N- and C-terminal residues of an assigned polypeptide stretch are retained during the course of resonance assignments. This results in faster execution time of the program, increased efficiency and greater robustness towards missing peaks.

SEQUENCE-specific resonance assignment (hereafter abbreviated as ssr_assignments) of all the NMR-active nuclei in a protein constitutes the first step towards the complete characterization of its three-dimensional structure[1]. Ssr_assignments, if carried out manually, constitute a tedious and time-consuming task. For this reason, there have been several efforts to automatize the assignment process[2]. With the advent of multidimensional, triple-resonance (^{1}H, ^{13}C, ^{15}N) strategies for resonance assignments, it has become increasingly clear that the information content of protein spectra can allow complete automation of ssr_assignments[3]. This advance has tremendous implications for the growth of NMR spectroscopy as a powerful tool in structural genomics[4,5], which involves high-throughput protein structure determination.

A number of strategies have been proposed for ssr_assignments in proteins[2]. These include approaches which utilize information from triple-resonance experiments and methods such as simulated annealing[6], Bayesian statistics and artificial intelligence[7], characteristic $^{13}C^{a}$ and $^{13}C^{b}$ chemical shifts of individual residues[8], threshold-accepting algorithm[9], connectivity-tracing algorithms[10] and neural networks[11].

We have recently proposed a novel algorithm for such automated assignments called TATAPRO (**T**racked **A**utomated **A**ssignments in **Pro**teins)[12]. TATAPRO achieves ssr_assignments of $^{1}H^{N}$, $^{13}C^{a}$, $^{13}C^{b}$, $^{13}C'/^{1}H^{a}$ and ^{15}N spins by utilizing the protein primary sequence and peak lists from a set of 3D triple-resonance spectra[12], namely CBCANH[13], CBCA(CO)NH[14], HNCO[15] and HN(CA)CO[16]. Peak lists obtained from these spectra consisting of chemical shift coordinates of the peaks, $(w_1, w_2, w_3) = (^{13}C/^{1}H^{a}, ^{15}N, ^{1}H^{N})$, along with their intensities and phases are taken as inputs for TATAPRO. For example, automatically picked CBCA(CO)NH peak list has information about $^{13}C^{a}_{i-1}$ and $^{13}C^{b}_{i-1}$ chemical shifts for a given pair of $^{15}N_i$ and $^{1}H^{N}_i$. From such a list, the chemical shifts

*For correspondence. (e-mail: chary@tifr.res.in)

of $^{13}C^{a}_{i-1}$ and $^{13}C^{b}_{i-1}$ are identified for each specific pair of $^{15}N_i$ and $^{1}H^{N}_{i}$ chemical shifts within the user-defined tolerance limits and grouped into a single set. Likewise, automatically picked CBCANH peak list has information about $^{13}C^{a}_{i}$, $^{13}C^{b}_{i}$, $^{13}C^{a}_{i-1}$ and $^{13}C^{b}_{i-1}$ chemical shifts, for a given pair of $^{15}N_i$ and $^{1}H^{N}_{i}$ chemical shifts. On the other hand, automatically picked HN(CA)CO and HNCO peak lists provide information about $^{13}C'_{i}$ and/or $^{13}C'_{i-1}$ chemical shifts. Besides, the success of TATAPRO relies on the classification of the 20 amino acid residues into eight different categories, which is based on their characteristic $^{13}C^{a}$ and $^{13}C^{b}$ chemical shifts, derived statistically from a large database of chemical shifts in the BioMagResBank (BMRB)[17]. The program has been successfully tested for resonance assignments using experimental data in three different proteins in the molecular weight range of 15–20 kDa (ref. 12). The program has also been tested for its robustness using published assignment data of four other proteins in the molecular weight range of 18–42 kDa (ref. 12).

During the course of its application to various different proteins, it was felt that the efficiency and speed of TATAPRO could be further improved by incorporating several changes in the program. The software has been modified accordingly, which has resulted in its increased robustness towards spectral overlap and missing data, a feature that is common when dealing with large molecular weight proteins. In this communication, we provide a detailed description of the improved version TATAPRO II and the results obtained thereby. The improved version of the program, called TATAPRO II, is available on request from chary@tifr.res.in or through the BMRB web server: www.bmrb.wisc.edu.

An extensive statistical analysis was carried out using the $^{13}C^{a}$ and $^{13}C^{b}$ chemical shift information of all proteins currently deposited in the BMRB. Out of 1832 proteins for which chemical shifts are available, ^{13}C shifts are available for 839 proteins in the molecular weight range of 10–42 kDa. The total number of $^{13}C^{a}$ and $^{13}C^{b}$ chemical shifts reported so far is 58,768 and 45,463 respectively. The total number of amino acid residues for which both $^{13}C^{a}$ and $^{13}C^{b}$ shifts are available is 42,629. The distribution of chemical shifts for individual amino acid residues is given in Table 1.

Figure 1 shows a 2D plot of $^{13}C^{a}$ and $^{13}C^{b}$ chemical shifts. As is evident from the figure, one can classify amino acid residues into eight distinct categories (as against seven done earlier[12]) based entirely on the characteristic $^{13}C^{a}$ and $^{13}C^{b}$ chemical shifts (each of these categories is shown with a different colour): (i) Gly having no $^{13}C^{b}$; (ii) Ala having $14 < {}^{13}C^{b} < 24$ ppm; (iii) Arg, Cys^{red}, Gln, Glu, His, Lys, Met, Val, Trp having $^{13}C^{b}$ in the range 24–36 ppm; (iv) Asp, Asn, Cys^{oxd}, Ile, Leu, Phe and Tyr having $^{13}C^{b}$ in the range 36–52 ppm; (v) Ser having $^{13}C^{b}$ in the range 56–67 ppm; (vi) Thr having $^{13}C^{b} > 67$ ppm; (vii) some Val residues having $^{13}C^{a} > 64$ ppm and $24 < {}^{13}C^{b} < 36$ ppm, and (viii) some Ile residues having $^{13}C^{a} > 64$ ppm and $36 < {}^{13}C^{b} < 52$ ppm. Pro residues, which lack an amide proton ($^{1}H^{N}$), are not observed in any of the aforementioned triple resonance spectra and hence are classified as the ninth category. In our algorithm, each of these nine categories is distinguished from one another

Table 1. Distribution of $^{13}C^{a}$ and $^{13}C^{b}$ chemical shifts for individual amino acid residues used in the statistical analysis. Pro residues are excluded from the analysis as they do not show up in any of the triple-resonance spectra considered in the present algorithm (see text)

	Number of chemical shifts		
Amino acid	$^{13}C^{a}$	$^{13}C^{b}$	$^{13}C^{a}$ and $^{13}C^{b}$
Alanine	4617	4006	3790
Arginine	2935	2379	2102
Asparagine	2590	2181	1978
Aspartic acid	3832	3275	2978
$Cysteine^{oxd}$	460	334	299
$Cysteine^{red}$	572	506	490
Glutamic acid	2561	2120	2019
Glutamine	4834	4092	3891
Glycine	4396	0	0
Histidine	1305	1085	997
Isoleucine	3284	2787	2553
Leucine	5216	4336	4213
Lysine	4651	3890	3598
Methionine	1355	1090	1009
Phenylalanine	2319	1950	1887
Serine	3669	2976	2812
Threonine	3382	2761	2689
Tryptophan	686	580	456
Tyrosine	1939	1592	1456
Valine	4165	3523	3412
Total	58,768	45,463	42,629

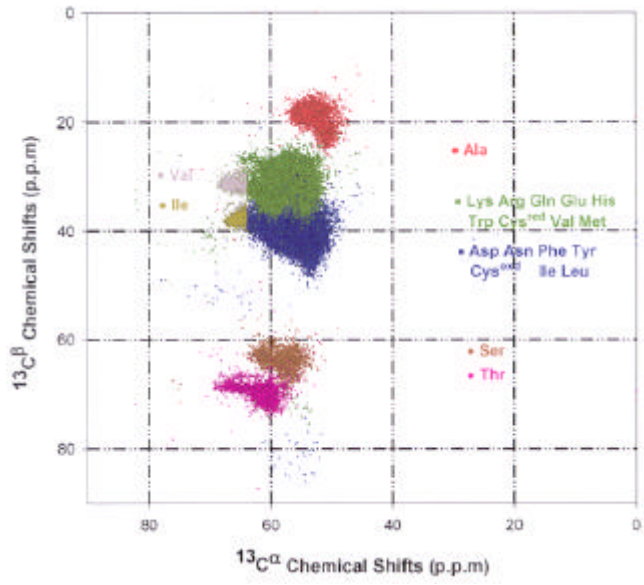


Figure 1. Two-dimensional plot of 42,629 $^{13}C^{a}$ vs $^{13}C^{b}$ chemical shifts obtained from 839 proteins. The chemical shifts have been grouped into 7 categories, each shown with a different colour, based on the characteristic $^{13}C^{a}$ vs $^{13}C^{b}$ shifts (see text).

by a two-digit code. In the earlier version of TATAPRO, amino acid residues were classified into eight different groups (Pro residues as the eighth category), with both Ser and Thr put under the same category[12]. Their separation into separate categories in the current version of the program has resulted in accomplishing the assignments in relatively less number of iterations and with an increased percentage of assignments as discussed below.

Table 2 shows the different categories of amino acid residues described above. The last two columns in Table 2 indicate percentages of residues that violate the two-digit code assigned to them. This happens only if a given residue exhibits unusual $^{13}C^{a}$ or/and $^{13}C^{b}$ chemical shift(s) as a result of which, it acquires a code different from the one generally expected. As is evident from Table 2, no dramatic differences have been observed in the percentage of $^{13}C^{b}$ chemical shift violations with the increased amount of chemical shift information, which is presently available in the BMRB. Interestingly, most of the violations are relatively less compared to what has been found earlier[12] (shown in parenthesis). The only exceptions are Ser and Thr, as they are presently given different two-digit codes. This is primarily due to the partial overlap of their $^{13}C^{b}$ chemical shifts. However, such an overlap does not have any adverse effect on our assignment procedure. On the contrary, as discussed below, the algorithm becomes more efficient and robust towards missing peaks.

Ssr_assignments using TATAPRO II can be divided into four essential steps: (i) preparing the chemical shift list (*master_list*). This list contains individual sets or rows of $^{1}H^{N}_{i}$, $^{15}N_{i}$, $^{13}C^{a}_{i}$, $^{13}C^{b}_{i}$, $^{13}C'_{i}$, $^{13}C^{a}_{i-1}$, $^{13}C^{b}_{i-1}$, and $^{13}C'_{i-1}$, chemical shifts; (ii) assigning two-digit codes to the individual rows in the *master_list* using the criteria shown in Table 2; (iii) starting assignments; this results in the creation of an *assig_array* and *pps_array* (described below); and (iv) mapping of the *assig_array* onto the *pps_array* for final ssr_assignments.

To begin with, the program reads in the $^{13}C^{a}_{i}$, $^{13}C^{b}_{i}$, and $^{13}C'_{i}$ chemical shift values from the first row in the *master_list* and searches for a row where, within the user-defined tolerance limits, these three shifts are seen as $^{13}C^{a}_{i-1}$, $^{13}C^{b}_{i-1}$ and $^{13}C'_{i-1}$ chemical shifts. If the search is successful and unique, the two-digit code associated with the new row is stored in an *assig_array*. This procedure corresponds to forward assignment in the primary sequence, which is continued till a break is encountered. The break can be due to a Pro residue, a missing peak(s) or if the C-terminal end of the polypeptide chain has been reached. Once a stretch of amino acid residues has been assigned in the forward direction, the program continues with the assignment in the backward direction, starting again from the first row in the *master_list*. For backward assignment, the program reads in the $^{13}C^{a}_{i-1}$, $^{13}C^{b}_{i-1}$, and $^{13}C'_{i-1}$ chemical shifts for a given row in the *master_list*, and searches for the row where these chemical shifts are seen as $^{13}C^{a}_{i}$, $^{13}C^{b}_{i}$ and $^{13}C'_{i}$ chemical shifts. If the search is successful and unique, the two-digit code associated with the new row is stored in the same *assig_array*, as was done in the case of forward assignment. The assignment is continued till a break is encountered. During this procedure, if more than one possible pair of $^{1}H^{N}_{i}$, and $^{15}N_{i}$ chemical shifts satisfies the assignment condition, the program continues with the assignment along each possible pathway until a break is encountered. After assigning a stretch of amino acid residues in both forward and backward directions, the program maps the *assig_array(s)* thus obtained onto the *pps_array* as the final step of ssr_assignments.

In the older version of TATAPRO, once a stretch of amino acid residues was assigned and before commencing with the next round of assignments, assigned rows and residues were deleted from the *master_list* and *pps_array* respectively. However, this resulted in the loss of crucial information, particularly for residues which constitute the N- and C-terminal ends of the assigned

Table 2. Two-digit code assigned to different amino acid residues based on their characteristic $^{13}C^{a}$ and $^{13}C^{b}$ chemical shift ranges obtained using the most recent database of chemical shifts in the BMRB

$^{13}C^{a}$ and $^{13}C^{b}$ chemical shift (d in ppm) characteristics	Amino acid	Two-digit code	Percentage of $^{13}C^{b}$ chemical shift violations*	Percentage of other residues taking the code*
Absence of $^{13}C^{b}$	Gly	1 0	0.0 (0.0)	0.0 (0.0)
$14 < d\ (^{13}C^{b}) < 24$	Ala	2 0	2.29 (0.8)	0.09 (0.09)
$56 < d\ (^{13}C^{b}) < 67$	Ser	3 0	3.06 (0.5)	0.38 (0.04)
$24 < d\ (^{13}C^{b}) < 36$ and $d\ (^{13}C^{a}) < 64$	Lys, Arg, Gln, Glu, His, Trp, Cys^{red}, Val and Met	4 0	3.16 (3.1)	2.59 (2.4)
$24 < d\ (^{13}C^{b}) < 36$ and $d\ (^{13}C^{a}) \geq 64$	Val	4 1	5.02 (1.3)	2.57 (0.6)
$36 < d\ (^{13}C^{b}) < 52$ and $d\ (^{13}C^{a}) < 64$	Asp, Asn, Phe, Tyr, Cys^{oxd}, Ile and Leu	5 0	3.16 (3.0)	4.20 (2.4)
$36 < d\ (^{13}C^{b}) < 52$ and $d\ (^{13}C^{a}) \geq 64$	Ile	5 1	8.04 (6.8)	1.85 (0.3)
–	Pro	6 0	–	–
$d\ (^{13}C^{b}) > 67$	Thr	7 0	5.25 (0.5)	0.37 (0.04)

*Numbers in parentheses indicate the percentage of violations obtained with the older version of TATAPRO.

Table 3. Details of test proteins and percentage of assignments obtained in each case using the old and the new versions of TATAPRO

Protein	Molecular weight (kDa)	Percentage of assignments obtained on random deletion of peaks*		
		0	15	30
Drosophila numb phosphotyrosine-binding domain	17.8	100	81.9 (80.1)	67.1 (54.5)
Fibroblast collagenase	18.7	100	71.3 (70.6)	61.4 (55.5)
Borrelia burgdorferi OspA	28	100	80.6 (80.2)	60.6 (58.5)
Escherichia coli maltose-binding protein	42	100	71.6 (70.7)	61.4 (57.2)

*Numbers in parentheses indicate the percentage of assignments obtained in test proteins using the older version of TATAPRO.

polypeptide stretch. These residues, if retained in the *master_list* and *pps_array*, could be used as markers. Hence, they help to assign those polypeptide stretches which lie in the immediate neighbourhood (in their N- or C-terminal sides) of the already assigned polypeptide stretches. Moreover, end-residues also serve as markers in resolving the ambiguity arising when the *assig_array* maps to more than one stretch in the *pps_array*. In light of this, in TATAPRO II, the N- and C-terminal residues of an assigned polypeptide stretch are retained in both the *master_list* and the *pps_array*. However, these amino acid residues are given a different two-digit code to distinguish them from those that are yet to be assigned. For example, an Ala that has been assigned and simultaneously lies either at the N- or the C-terminus of an assigned polypeptide stretch, is given a new code **2 1** in the *master_list* and *pps_array*. This distinguishes it from other unassigned Ala residues in the *master_list* and *pps_array* that have a code **2 0** (see Table 2). Such inclusion of the end-residues resulted in improved efficiency of the program. The program in the present form requires less number of iterations to find the correct *assig_array*, resulting in an increased speed of execution.

Table 3 summarizes the results obtained for ssr_assignments in the test proteins using TATAPRO II. With no random deletion of rows from the *master_list*, 100% assignments are obtained using both old and new versions of the program. This is because the initial data set is obtained using the published assignment of all the amino acid residues, and hence is perfect. However, the robustness of the program is tested on random deletion of rows from the *master_list*, which mimics a real situation wherein one encounters missing peaks. Even after random deletion of rows up to 30% from the *master_list*, TATAPRO II gives increased percentage of assignments in all the test proteins compared to the older version of the program.

Another modification concerns residues in an assigned polypeptide stretch that have unusual chemical shifts and therefore do not belong to their respective category. In such a situation (which we refer to as a 'mismatch'), the mapping of *assig_array* onto the *pps_array* will result in either an incorrect mapping or no mapping. In the older version of the program, all residues were equally considered as candidates for a mismatch. However, refined statistics carried out using the latest database of chemical shifts reveals that Gly, Ala, Ser and Thr have the least $^{13}C^{a}$ and $^{13}C^{b}$ chemical shift violations (Table 2). Hence, these residues rarely have mismatches. In light of this, in TATAPRO II, these residues are not considered to be mismatches during any stage of assignments. This substantially reduces the errors which occur if these residues are considered as mismatches. Further, during all stages of the assignments in TATAPRO II, a maximum of only one mismatch is allowed for mapping *assig_array* onto the *pps_array*. This is unlike the older version, wherein a maximum of three mismatches were allowed. This further decreased the chances of incorrect mapping. However, such criterion did not influence the final percentage of assignments obtained.

In conclusion, the modifications incorporated in TATAPRO II include refinement of the $^{13}C^{a}$ and $^{13}C^{b}$ chemical shift statistics and retaining the N- and C-terminal residues in an assigned polypeptide stretch in both the *master_list* and *pps_array*. This results in less number of iterations during the entire assignment procedure and hence, faster execution time of the program and concomitant increased efficiency in ssr_assignments. TATAPRO II has been tested using published data on four proteins in molecular weight range of 18–42 kDa, and shows increased percentage of assignments compared to the earlier version of TATAPRO.

1. Wüthrich, K., *NMR of Proteins and Nucleic Acids*, John Wiley, New York, 1986.
2. Moseley, H. N. B. and Montelione, G. T., *Curr. Opin. Struct. Biol.*, 1999, **9**, 635–642.
3. Ikura, M., Kay, L. E. and Bax, A., *Biochemistry*, 1990, **29**, 4659–4667.
4. Montelione, G. T., Zheng, D., Huang, Y. J., Gunsalus, K. C. and Szyperski, T., *Nature Struct. Biol.* (*Suppl.*), 2000, **7**, 982–985.

5. Chary, K. V. R. and Atreya, H. S., *J. Postgrad. Med*., 2002, **48**, 83–87.
6. Buchler, N. E. G., Zuiderwig, E. R. P., Wang, H. and Goldstein, R. A., *J. Magn. Reson*., 1997, **125**, 34–42.
7. Zimmerman, D. E. *et al*., *J. Mol. Biol.*, 1997, **269**, 592–610.
8. Friedrichs, M. S., Mueller, L. and Wittekind, M., *J. Biomol. NMR*, 1994, **4**, 703–726.
9. Leutner, M., Gschwind, R. M., Liermann, J., Schwarz, C., Gemmecker, G. and Kessler, H., *ibid*, 1998, **11**, 31–43.
10. Olson, J. B. Jr. and Markley, J. L., *ibid*, 1994, **4**, 385–410.
11. Hare, B. J. and Prestegard, H., *ibid*, 1994, **4**, 35–46.
12. Atreya, H. S., Sahu, S. C., Chary, K. V. R. and Govil, G., *ibid*, 2000, **17**, 125–136.
13. Wittekand, M. and Mueller, L., *J. Magn. Reson.*, 1993, **B101**, 201–205.
14. Grzesiek, S. and Bax, A., *J. Am. Chem. Soc*., 1992, **114**, 6291–6293.
15. Kay, L. E., Ikura, M., Tschudin, R. and Bax, A., *J. Magn. Reson*., 1990, **89**, 496–514.
16. Clubb, R. T., Thanabal, V. and Wagner, G., *ibid*, 1992a, **97**, 213–217.
17. Seavey, B. R., Farr, E. A., Westler, W. M. and Markley, J. L., *J. Biomol. NMR*, 1991, **1**, 217–236.

ACKNOWLEDGEMENTS. The facilities provided by the National Facility for High Field NMR, supported by Department of Science and Technology, Department of Biotechnology, Council of Scientific and Industrial Research, and Tata Institute of Fundamental Research, Mumbai are gratefully acknowledged. The help provided by two summer students, Mr Ajay Sheth and Mr Vikas Yadav, in the statistical analysis of $^{13}C^{a}$ and $^{13}C^{b}$ chemical shift information, is gratefully acknowledged.

Received 19 June 2002; revised accepted 16 October 2002

Mechanism of artificial transformation of *E. coli* with plasmid DNA – Clues from the influence of ethanol

Suchitra Sarkar, Sujan Chaudhuri and Tarakdas Basu*

Department of Biochemistry and Biophysics, University of Kalyani, Kalyani 741 235, India

The standard method of transformation of *E. coli* with plasmid DNA involves two important steps – binding of DNA to the cell surface, suspended in 100 mM $CaCl_2$ at 0°C, and the subsequent entry of DNA to the cell cytosol by a heat-pulse from 0 to 42°C. When competent *E. coli* cells were transformed with plasmid DNA in the presence of different concentrations (up to 10% v/v) of ethanol, the transformation efficiency $(TR)_E$ decreased gradually with increase in ethanol concentration. This decrease in $(TR)_E$ was directly proportional to ethanol-mediated leaching of lipopolysaccharide (LPS) molecules from the competent cell surface, indicating LPS was the major target site for DNA adsorption to the competent cells. *In vitro* spectrophotometric study showed evidence that there was binding interaction between plasmid DNA and *E. coli* LPS in the presence of a divalent cation, Ca^{2+}. Moreover, plasmid DNA, previously incubated with LPS in $CaCl_2$, had less ability to transform *E. coli* cells. The results suggest that during artificial transformation of *E. coli*, the naked DNA was first bound to the LPS molecules on the competent cell surface and uptake of this LPS-absorbed DNA into the cell cytosol was associated with $CaCl_2$-mediated cell-membrane disintegration.

THE technique of DNA transformation has become important in virtually all aspects of molecular genetics. Transformation is defined as the uptake and expression of foreign DNA by cells. Bacterial transformation occurs naturally in many species such as *Micrococcus*, *Haemophilus* and *Bacillus*[1,2]; all these organisms have proteins on their exterior surface whose function is to bind to DNA in their environment and transport it into the cell. However, it is still a rare event for most bacteria to naturally take up DNA from the environment. But by subjecting bacteria to certain artificial conditions, many of them are able to take up free DNA[3], and the cells in such state are referred to as competent.

In *E. coli* the competence can be developed by suspending the cells in ice-cold $CaCl_2$ and then subjecting to a brief heat-shock at 42°C (refs 4, 5). Although *E. coli* has developed into a universal host organism both for molecular cloning and for a diverse set of assays involving cloned genes, the technique of *E. coli* transformation is highly inefficient even using competent cells. The vast majority of DNA molecules added will not enter any cell, and the vast majority of bacterial cells will receive no DNA. Besides Ca^{2+} ions, other frequently used cations include Mg^{2+}, Mn^{2+}, Rb^{+}, for competence generation[3]. However, the exact mechanism by which DNA adsorbs to the *E. coli* cell surface and enters the cell cytosol, and why the transformation is stimulated by these treatments, is still largely obscure.

One proposed hypothesis[6] is that DNA crosses through the least-barrier path at zones of adhesion, where the outer and inner cell membranes fuse to pores in the cell wall. The zones of adhesion are rich in negatively-charged lipo polysaccharide (LPS) molecules[7,8] and DNA also being negatively charged, cannot enter the cell easily as the two negative polarities repel each other. A divalent cation such as calcium, is believed to form stable coordination complexes with phosphates, and thus may facilitate the association of the two phosphate-rich structures like DNA and LPS.

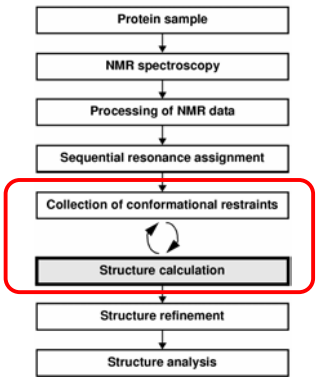
The transformation process, being a membrane-bound phenomenon, is most likely to be influenced by the well-

*For correspondence. (e-mail: tbasu@cal3.vsnl.net.in)

Automated NMR Structure Calculation

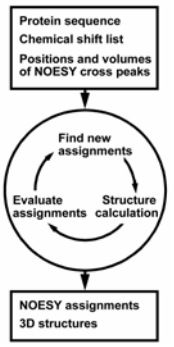
Peter Güntert
RIKEN Genomic Sciences Center
Yokohama

NMR structure determination



Automated NOESY assignment and structure calculation

- Automated methods are
  - much faster
  - more objective
- Problems may arise because of
  - imperfect input data
  - limitations of the algorithms used
- Iterative process: All but the first cycle use the structure from the preceding cycle.
- The first cycle is important for the reliability of the method.



Input data for structure calculation with automated NOESY assignment

- Amino acid sequence
- Residue library
- Chemical shift list(s)
- Peak list(s)
- Additional conformational constraints

Additional conformational constraints

- Torsion angle constraints
  - from C^{α} chemical shifts
  - from grid search
  - from external source, e.g. TALOS
  - for favored sidechain rotamer positions
  - for favored Ramachandran plot regions
- ^{3}J scalar coupling constants
- Hydrogen bond distance constraints
- …

Chemical shift list(s)

- XEASY (DYANA, CYANA) format:

	Shift	Error	Atom		
75	122.122	0.000	N	53	
76	8.235	0.000	HN	53	
77	54.245	0.000	CA	53	
78	4.352	0.000	HA	53	
79	41.823	0.000	CB	53	Arbitrary stereospecific assignment
80	1.510	0.030	HB2	53	
81	1.582	0.030	HB3	53	
83	26.664	0.000	CG	53	
84	1.472	0.000	HG	53	
85	0.647	0.000	QD1	53	
86	0.530	0.000	QD2	53	
87	24.011	0.000	CD1	53	Pseudo atom Q… for
91	22.839	0.000	CD2	53	degenerate shifts, e.g. methyls

- BMRB (BioMagResBank) format

Degenerate ^{1}H chemical shifts

- A group of degenerate protons is represented in the chemical shift list by a "pseudo atom" (or the corresponding "ambiguity index" in BMRB files)
- CYANA expands distance constraints to pseudo atoms into ambiguous distance constraints with all the corresponding protons represented by the pseudo atom(s): "$1/r^6$-summation". (Fletcher et al., *J. Biomol. NMR* **8**, 292 (1996))
- There are no "pseudo atom corrections" for upper distance bounds.

Pseudo atoms

- Degenerate pairs of methylene protons:
  `QB` ($H^{\beta 2}/H^{\beta 3}$), …
- Methyl groups:
  `QB` (Ala), `QG1`/`QG2` (Val), `QD1`/`QD2` (Leu), …
- Degenerate pairs of methyl groups:
  `QQG` (Val), `QQD` (Leu)
- Phe/Tyr aromatic ring protons:
  `QD` ($H^{\delta 1}/H^{\delta 2}$), `QE`, ($H^{\varepsilon 1}/H^{\varepsilon 2}$), `QR` (all ring protons)

Diastereotopic protons

- Stereospecifically assigned:
  - 2 entries in the chemical shift list, e.g.: `HB2`/`HB3`
  - CYANA command to supress swapping:
    `atom stereo "HB2 23"`
- Not stereospecifically assigned, not degenerate:
  - 2 entries in the chemical shift list: `HB2`/`HB3`
  - During the calculation CYANA will periodically check for the optimal stereo-assignment and swap the two protons, if needed. (Folmer et al., *J. Biomol. NMR* **9**, 245 (1997))
- Degenerate:
  - 1 pseudo atom entry in the chemical shift list: `QB`

Using foreign nomenclature

- CYANA commands:

```
translate bmrb
```
Invoke translation table between internal and foreign atom names

```
read prot demo.prot unknown=warn
```
Read chemical shift ("proton" list) — Skip unknown atoms

```
translate off
```
Return to standard CYANA nomenclature

Peak list(s)

- 2D, 3D, 4D NOESY
- XEASY format

h: 1st dimension is "free" ^{1}H dimension
C: 2nd dimension is ^{13}C dimension
H: 3rd dimension is ^{1}H bound to ^{13}C

```
# Number of dimensions 3
#FORMAT xeasy3D
#CYANAFORMAT hCH
1 3.893 42.959 3.893 1 U 8.76E+6 0.00E+0 a 0  0  0  0
2 4.165 62.033 4.165 1 U 1.81E+6 0.00E+0 a 0  0  0  0
3 4.079 61.908 4.079 1 U 7.77E+5 0.00E+0 a 0  0  0  0
```

··· Peak position (ppm) — Volume — Assignment (unassigned)

- NMRView format
- ANSIG format

Peak and chemical shift list(s)

- Normal case:
  - several NOESY peak lists (e.g. ^{15}N-resolved NOESY, ^{13}C-resolved NOESY, aromatic ^{13}C-resolved NOESY)
  - 1 chemical shift list
- Significant shift deviations between different NOESY spectra:
  - several NOESY peak lists
  - separate chemical shifts list for each NOESY peak list
- Agreement between peak positions and corresponding entries in the chemical shift list:
  e.g. ±0.03 ppm for ^{1}H, ±0.5 ppm for ^{13}C/^{15}N

Structure calculation with automated NOESY assignment
(CYANA 2.0)

```
peaks       := noeN,noeC          # peak lists
prot        := enth               # chemical shift list(s)
tolerance   := 0.025, 0.02, 0.4   # shift tolerances (ppm)
                                  # 1H(a), 1H(b), 13C/15N(b)
#cal        := 3.0E7, 7.0E8       # calibration constants
constraints :=                    # additional constraints

noeassign peaks=$peaks prot=$prot
```

Keep existing assignments

- To preserve part or all of the assignments in the input peak lists:
  - Define the subset of peaks for which assignments will be preserved (e.g. peaks with numbers 200…300):

```
subroutine KEEP
  peaks select "*, * number=200..300"
end
```

  - Call the **noeassign** command with the **keep** option:

```
noeassign peaks=$peaks prot=$prot \
          keep=KEEP
```

Running CYANA

- CYANA commands to execute a calulation are saved in a "macro", a script file with extension ".cya", e.g. `CALC.cya`
- Interactive mode:

```
cyana> CALC
```

- Parallel calculation on 20 CPUs:

```
% cyanajob -n 20 CALC
```

  - Output will be saved in a file `CALC.out`.
  - The `cyanajob` Unix shell script might need to be adapted for a particular computer or batch system.

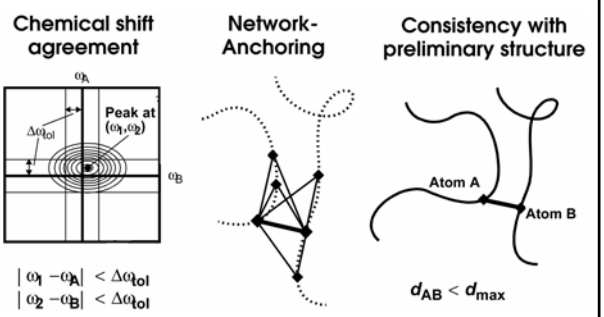
Automated NOE Assignment and Structure Calculation

- Ambiguous distance constraints
- Network-anchored assignment
- Constraint combination/ violation confinement

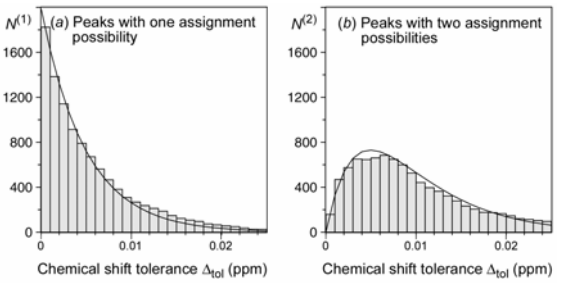
Assignment/structure calculation cycles
(CYANA 2.0)

- Cycle 1:
  - Automated NOE assignment based on chemical shift agreement and network-anchoring
  - Torsion angle dynamics structure calculation of 100 conformers
- Cycles 2-7:
  - Automated NOE assignment based on chemical shift agreement, 20 best conformers from previous cycle, and network-anchoring
  - Torsion angle dynamics structure calculation of 100 conformers
- Final structure calculation:
  - Using NOE assignments from cycle 7
  - Unambiguous constraints (split constraints with multiple contributions)
  - Stereospecific assignments, where possible
  - Torsion angle dynamics structure calculation of 100 conformers

Conditions for valid NOESY assignments



Chemical-shift based NOE assignment



NOE assignment probability

(CYANA 2.0)

Prob(*assignment to atoms A-B is correct*) =

Prob(*chemical shifts match*) x
Prob(*distance A-B < upper limit*) x
Prob(*other assignments predict NOE A-B*)

$$P_{tot} = P_{shift} \cdot P_{structure} \cdot P_{network}$$

Accept assignments with $P_{tot} > P_{min}$ (= 20%)

Chemical shift-based assignment probability

$$P_{shift} = \exp\left[-\frac{1}{2} \sum_{k=1}^{D} \left(\frac{\omega_k - \omega(A_k)}{\Gamma \Delta\omega_k} \right)^2 \right]$$

Dimensionality of peak (D = 2,3, or 4)
Peak position in dimension k
Chemical shift of atom A_k
weighting factor
Chemical shift tolerance in dimension k

Structure-based assignment probability

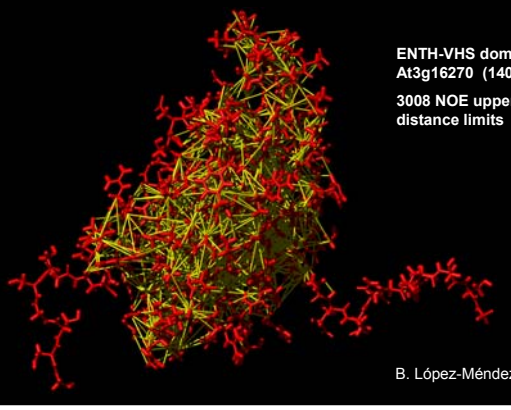
- Probability, derived from preliminary structures, that the distance d_{AB} is shorter than the upper distance limit u:

number of conformers in which $d_{AB} < u + \Delta u$
tolerance value that accounts for the limited precision of the preliminary structure

$$P_{structure} = \frac{N(d_{AB} < u + \Delta u)}{N}$$

total number of conformers that represent the preliminary structure

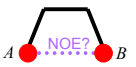
NOE network



ENTH-VHS domain
At3g16270 (140 a.a.)
3008 NOE upper distance limits

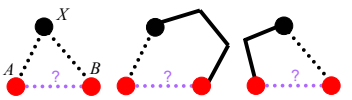
B. López-Méndez *et al.*

Network-anchoring

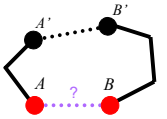


Distance d_{AB} restricted by covalent structure

Another NOE (e.g. a transposed peak) with the same assignment exists



A pair of NOEs (or an NOE and a covalently restricted distance) connecting atoms A and B through a third atom X exist



Atoms A and B are covalently close to two atoms A' and B' that are connected by a NOE

Individual contributions to network-anchoring probability

- A priori probability that two atoms in a protein of radius R are closer than the upper limit u: $P_1(d_{AB} \le u) = (u/R)^3$
- Covalent structure requires that $d_{AB} < u_c$: $P_2(d_{AB} \le u) = \min\left((u/u_c)^3, 1\right)$
- Another NOE (e.g., a transposed peak) with probability P' of having the same assignment and upper limit u' exists:

$$P_3(d_{AB} \le u) = P'_{tot}(d_{AB} \le u') \cdot \min\left((u/u')^3, 1\right)$$

- A pair of NOEs connecting atoms A and B through a third atom X exists:

$$P_4(d_{AB} \le u) = P_{tot}(d_{AX} \le u_{AX}) \cdot P_{tot}(d_{BX} \le u_{BX}) \cdot f(u; u_{AX}, u_{BX})$$

- Atoms A and B are covalently close to two atoms A' and B' that are connected by a NOE:

$$P_5(d_{AB} \le u) = P_{tot}(d_{A'B'} \le u_{A'B'}) \cdot f(u; u_{c,AA'}, u_{A'B'}, u_{c,BB'})$$

Network-anchoring-based assignment probability

$$P_{network} = 1 - (1 - P_1)(1 - P_2)\cdots$$

- $P_1, P_2, \ldots$ are probabilities that represent different possible ways to confirm that the assignment *A-B* corresponds to a short enough distance, i.e. that $d_{AB} < u$.
- $P_{network} \ge P_k$ for all $P_1, P_2, \ldots$

→ Network-anchoring requires that some (not all) of the individual probabilities $P_1, P_2, \ldots$ are high.

NOE assignment: Example

Peak number | used | Peak position | Upper distance bound

peak 2784 predicts this assignment with 76% probability

peak 2786 and a short covalent distance of 2.6 Å predict this assignment with 55% probability

Cycle 1 (no preliminary structure):

```
Peak 109 (6.77, 9.69, 128.31 ppm; 3.84 A):
2 out of 6 assignments used, quality = 0.98:
QE     TYR    83 + HE1    TRP    47  OK    91    97   -  93     -     2784=76, 2786/2.6=55
HZ3    TRP    47 + HE1    TRP    47  OK    82    97   -  84     -     4.3/131=43, 2.4/4=29
HZ     PHE    43 - HE1    TRP    47  lone  18    95   -  19     -     3.8/4=7, 132/2.8=5
HE22   GLN    32 - HE1    TRP    47  lone  17    94   -  19     -     132/2.8=5, 1592/2.6=4
HE22   GLN   106 - HE1    TRP    47  lone   3    98   -   3     -
HN     ASP-   79 - HE1    TRP    47  lone   1    61   -   2     -
```

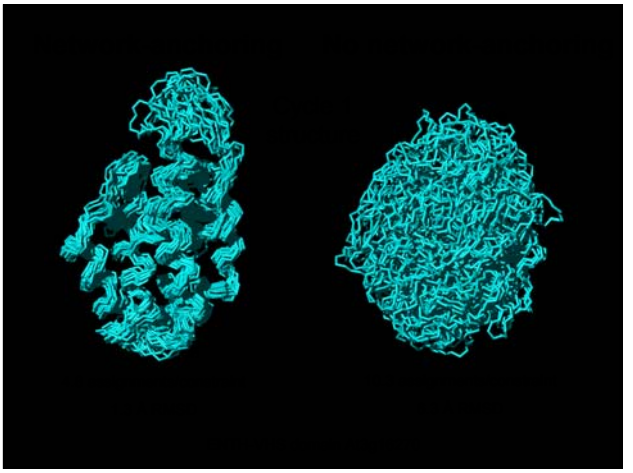
assignment possibility | not used, $P_{tot} < 20\%$ | $P_{tot} = P_{shift} \times P_{network}$

Cycle 3 (using preliminary structure from cycle 2):

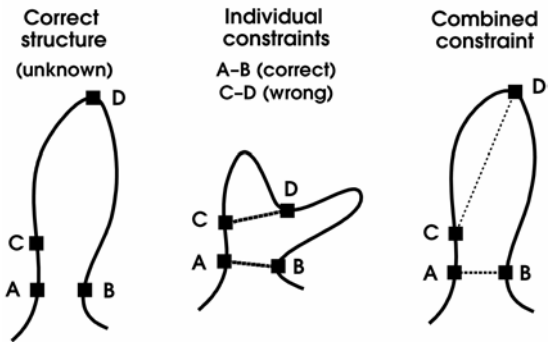
distance range in preliminary structure

```
Peak 109 (6.77, 9.69, 128.31 ppm; 3.84 A):
1 out of 2 assignments used, quality = 0.90:
QE     TYR    83 + HE1    TRP    47  OK    90    97 100  93  1.9-3.4  2784=76, 1592/2.6=60
HZ     PHE    43 - HE1    TRP    47  lone   1    95  60   1  3.2-11.1
Violated in 0 structures by 0.00 A.
```

$P_{structure}$



Constraint Combination



Constraint combination

- **Problem:** Peaks with wrong (long-range) assignments may severely distort the structure, especially in the first cycles of automated NOE assignment and structure calculation, and may lead to convergence to a wrong structure.
- **Idea:** From two long-range peaks each, combine the assignments into a single distance constraint.
  → Occurence of erroneous constraints is reduced.

Ambiguous distance constraints

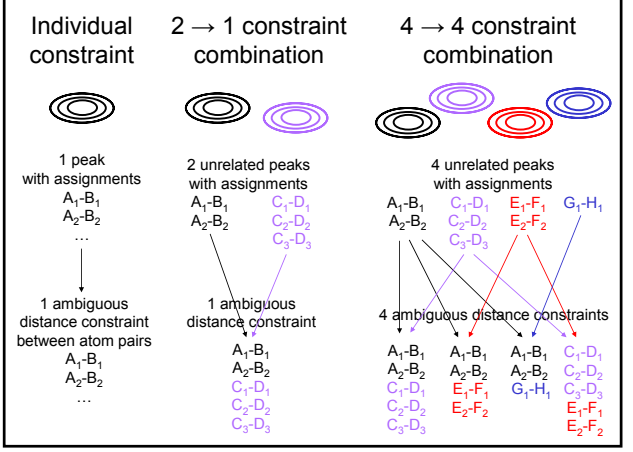
$$d_{\text{eff}} = \left(\sum_k d_k^{-6} \right)^{-1/6} \leq b$$

b: upper distance bound; d_k: distance for assignment possibility k; $\sum_k$: sum over all assignment possibilities

- Constraint with multiple assignments
- If one assignment possibility leads to a sufficiently short distance, then the ambiguous distance restraint will be fulfilled.

→ The presence of wrong assignment possibilities has no (or little) influence on the structure, as long as the correct assignment possibility is present.

Nilges et al., *J. Mol. Biol.* **269**, 408–422 (1997)



Effect of constraint combination

- **Example:** 1000 long-range peaks, 10% of which would lead to erroneous constraints.
- Individual constraints: 1000 constraints, ≈1000 x 0.1 = 100 wrong
- 2 → 1 constraint combination: 500 constraints, ≈500 x 0.1^2 = 5 wrong
- 4 → 4 constraint combination: 1000 constraints, ≈1000 x 0.1^2 = 10 wrong

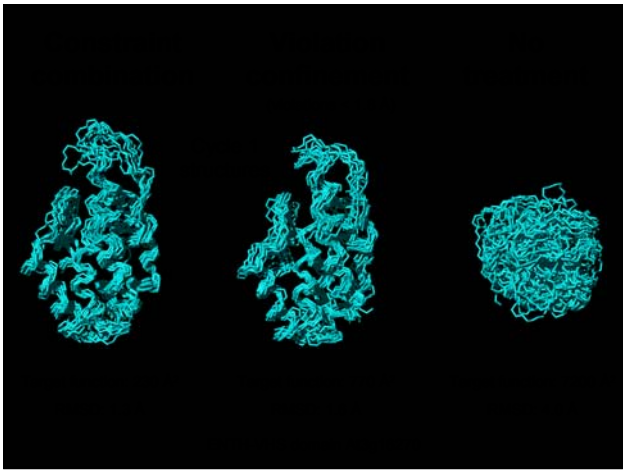
Violation confinement

- Confine the size of a distance constraint violation to a given maximal value v_{max}
- Ambiguous distance constraint with upper bound u:

$$d_{\text{eff}} = \left((u + v_{\text{max}})^{-6} + \sum_{k=1}^{n} d_k^{-6} \right)^{-1/6}$$

→ d_{eff} is always smaller than $u + v_{\text{max}}$

→ The contribution to the target function, $(d_{\text{eff}} - u)^2$, never exceeds v_{max}^2

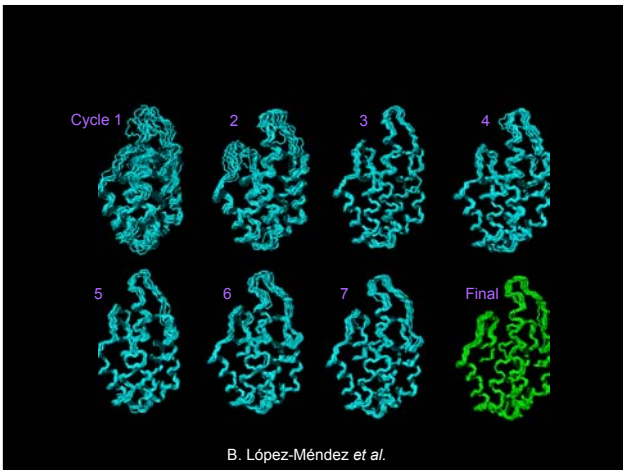


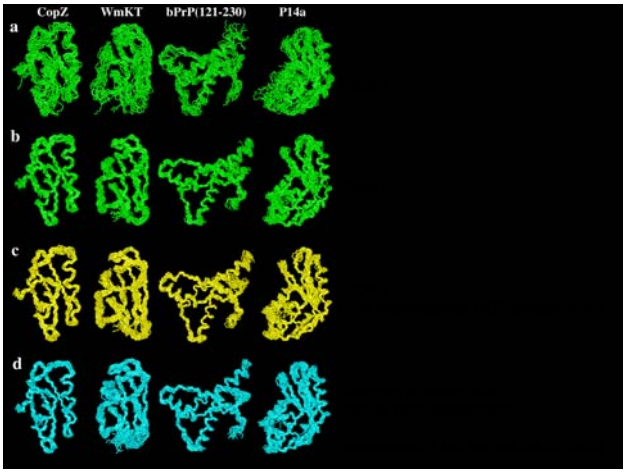
Schedule (CYANA 2.0)

Cycle	Violation tolerance	Stereo pairs	Calibration elasticity	Constraint combination	Ramach./ rotamers
1	-	swap	0%	yes	yes
2	1.5 Å	swap	0%	yes	yes
3	0.9 Å	swap	0%	no	yes
4	0.6 Å	swap	-0%/+25%	no	yes
5	0.3 Å	swap	-0%/+25%	no	yes
6	0.1 Å	swap	-0%/+25%	no	yes
7	0.1 Å	swap	-0%/+25%	no	yes
final	(cycle 7)	assign/ symmetrize	(cycle 7)	no	weak

Computation time

- Complete NOE assignment and structure calculation of the ENTH-VHS domain At3g16270 (140 a.a.) with CYANA (8 × 100 conformers; 10000 torsion angle dynamics steps per conformer)
- Linux, 7 processors, Pentium IV,1.8-3.06 GHz **130 min**
- SGI, 20 processors, 400 MHz IP35 **93 min**





Output files from a CYANA run

- From each cycle (*N* = 1,…,7):
  - `cycle`*N*`.noa` NOE assignment details for each peak
  - `cycle`*N*`.upl` NOE upper distance limits
  - `cycle`*N*`.cor` Bundle of conformers
  - `cycle`*N*`.ovw` Target function/violation overview
- In addition for the last cycle (cycle 7)
  - *peaklist*`-cycle7.peaks` NOESY peak lists with assignments from CYANA (name of input peak list: *peaklist*`.peaks`)
- From the final structure calculation:
  - `final.upl` Final NOE upper distance limits
  - `final.cor` Final bundle of conformers
  - `final.ovw` Target function/violation overview

Analyzing the output

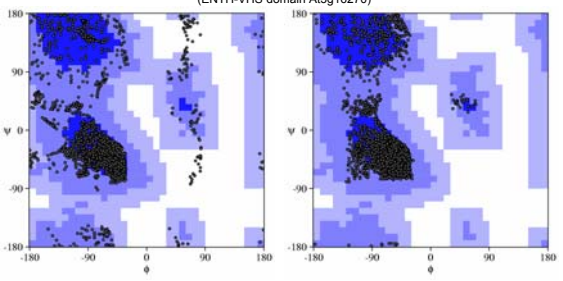
- `cyanatable` (Unix shell script): creates an overview of the NOE assignments and the results of the structure calculations of each cycle.
- `cyanalist` (Unix shell script): extracts information about certain classes of peaks (e.g. unassigned, violated) from the (typically large) `cycle`*N*`.noa` files.

Output overview table

Cycle	:	1	2	3	4	5	6	7	final
Peaks:									
selected	:	5439	5439	5439	5439	5439	5439	5439	
with assignment	:	5100	4806	4742	4749	4712	4678	4675	
without assignment	:	339	633	697	690	727	761	764	
with diagonal assignment	:	12	12	12	12	12	12	12	
Cross peaks:									
with off-diagonal assignment	:	5088	4794	4730	4737	4700	4666	4663	
with unique assignment	:	675	3591	3872	3950	4115	4195	4194	
with short-range assignment |i-j|<=1	:	3295	3208	3165	3154	3120	3102	3089	
with medium-range assignment 1<|i-j|<5	:	1020	925	921	914	904	884	893	
with long-range assignment |i-j|>=5	:	773	661	644	669	676	680	681	
Upper distance limits:									
total	:	3786	2996	2832	2789	2707	2643	2683	2731
short-range, |i-j|<=1	:	2007	1586	1486	1440	1388	1348	1273	1304
medium-range, 1<|i-j|<5	:	1220	959	787	775	751	726	760	765
long-range, |i-j|>=5	:	559	451	559	574	568	569	650	662
Average assignments/constraint	:	4.81	1.73	1.27	1.25	1.18	1.14	1.00	1.00
Average target function value	:	230.84	69.79	68.20	9.22	3.99	2.98	1.70	0.43
RMSD (residues 15..130):									
Average backbone RMSD to mean	:	1.34	0.97	0.57	0.67	0.68	0.60	0.53	0.53
Average heavy atom RMSD to mean	:	1.76	1.44	1.09	1.19	1.20	1.07	0.98	1.01

Ramachandran plot constraints

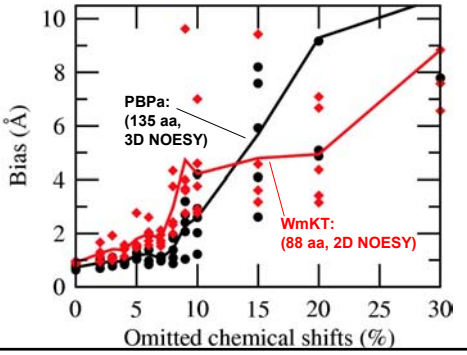
without (ENTH-VHS domain At3g16270) with



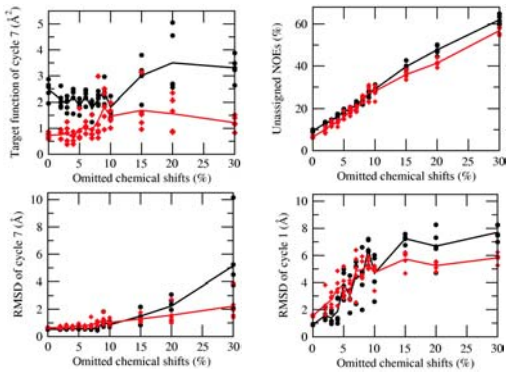
73% in most favored regions
21% in additionally allowed regions
4% in generously allowed regions
2% in disallowed regions

77% in most favored regions
23% in additionally allowed regions
0% in generously allowed regions
0% in disallowed regions

Effect of missing chemical shifts



How to recognize wrong structures?



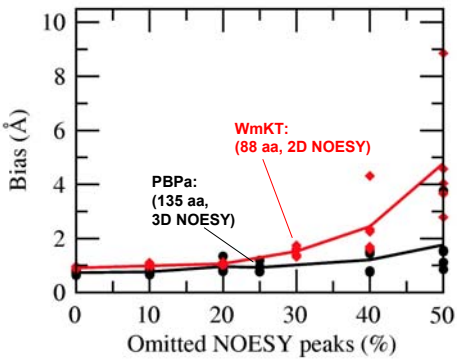
Has it worked?

- ~~Low final CYANA target function values:~~
  ~~Since peaks that are incompatible with the structure obtained in the previous cycle will remain unassigned, in general low target function values for the final structure are obtained with any input in cycle 7.~~
- ~~Small RMSD of final structure~~
  ~~Indicates only precision, not accuracy.~~

Has it really worked?

- Criteria for successful automated NOE assignment and structure calculation:
- 90% or more of non-labile and backbone ^{1}H chemical shifts assigned
- RMSD < 3 Å in cycle 1
  (average RMSD of the individual conformers to their mean coordinates for the backbone atoms N, Cα, C', excluding unstructured regions)
- Less than 20% of peaks with exclusively long-range assignments discarded

Effect of missing NOESY peaks



Troubleshooting

- Correct and complete chemical shift list
- Correct peak list:
  - Remove artifact peaks
  - Correct peak positions
  - Correct peak volumes
  - Add additional peaks
- Check agreement between peak positions and chemical shift positions
- Repeat calculation

Acknowledgments

Torsten Herrmann
Jun-Goo Jee
Blanca López-Méndez
David Pantoja-Uceda

ELSEVIER

Progress in Nuclear Magnetic Resonance Spectroscopy xx (2003) xxx–xxx

PROGRESS IN NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY

www.elsevier.com/locate/pnmrs

Automated NMR protein structure calculation

Peter Güntert*

RIKEN Genomic Sciences Center, 1-7-22 Suehiro, Tsurumi, Yokohama 230-0045, Japan

Accepted 23 June 2003

Contents

* Tel.: +81-45-503-9345; fax: +81-45-503-9343.
E-mail address: guentert@gsc.riken.go.jp (P. Güntert).

0079-6565/03/$ - see front matter © 2003 Published by Elsevier B.V.
doi:10.1016/S0079-6565(03)00021-9

Keywords: Protein structure; Chemical shift assignment; Conformational constraints; Automated structure determination; Automated assignment

1. Introduction

The NMR method for protein structure determination in solution is now firmly established besides X-ray crystallography as a second generally applicable technique that can give a detailed picture of the three-dimensional structure of biological macromolecules at atomic resolution. By April 2003, more than 3150 (15%) of the entries deposited in the Protein Data Bank [1] originated from macromolecular structures that had been solved by NMR methods. NMR plays also an important role in the current efforts of structural genomics that are driven by the vision to supplement the knowledge on the sequence of proteins by structural information on a genome-wide scale, determined either experimentally or by theoretical homology modeling [2]. Structural genomics wants to help us understand the molecular 'book of life', the genome, by translating its concise but cryptic DNA or amino acid sequence idiom into the more readily comprehensible language of three-dimensional structures. A massive structure determination effort will be needed to achieve the aim of structural genomics, since of the order of 10^5 new protein structures need to be determined experimentally [3] in order to allow coverage of the rest of sequence space with structures from theoretical methods because at present homology modeling is reliable only for proteins that share high (more than 30%) sequence identity with a protein of known three-dimensional structure.

Until recently NMR protein structure determination has remained a laborious undertaking that occupied a trained spectroscopist over several months for each new protein structure. It has been recognized that many of the time-consuming interactive steps carried out by an expert during the process of spectral analysis could be accomplished by automated, computational approaches [4]. Today automated methods for NMR structure determination are playing a more and more prominent role and will most likely supersede the conventional manual approaches to solving three-dimensional protein structures in solution.

This review gives an introduction to the current state of automated NMR structure calculation. Section 2 gives a general survey of the principles and problems of automated NOESY assignment and structure calculation. Section 3 is devoted to various specific implementations of algorithms for automated NOESY assignment and structure calculation. Aspects of reliability, quality control and troubleshooting in automated NMR structure calculation are discussed in Section 4. Alternative methods for structure calculation without chemical shift assignment are introduced in Section 5. In the three core Sections 3–5 a selection of programs is presented for which either the literature bears testimony of widespread use or that embody concepts of particular interest and future potential.

For consistency and simplicity, the following conventions are used in this review: an interaction between two or more nuclei is manifested by a *signal* in a multidimensional spectrum. A *peak* refers to an entry in a peak list that has been derived from an experimental spectrum by *peak picking*. A peak may or may not represent a signal, and there may be signals that are not represented by a peak. *Chemical shift assignment* is the process and the result of attributing a specific chemical shift value to a nucleus. *Peak assignment* is the process and the result of identifying

in each spectral dimension the nucleus or nuclei that are involved in the signal represented by the peak. *NOESY assignment* is peak assignment in NOESY spectra.

2. General principles of automated NOESY assignment and structure calculation

Many approaches have already been proposed in order to automate parts of the NMR protein structure determination process. So far, all *de novo* NMR protein structure determinations have followed the 'classic' way [5] including the successive steps of sample preparation, NMR experiments, spectrum calculation, peak picking, chemical-shift assignment, NOESY assignment and collection of other conformational constraints, structure calculation, and structure refinement. Alternative approaches that bypass the potentially cumbersome chemical shift and NOESY assignment steps have been proposed, and will be discussed in Section 5 below. The present section introduces basic aspects of automated NOESY assignment that are relevant for any algorithm implementing the standard approach.

2.1. Chemical shift assignment

The assignment of NOESY cross peaks requires as a prerequisite a knowledge of the chemical shifts of the spins from which nuclear Overhauser effects (NOEs) are arising. There have been many attempts to automate this chemical shift assignment step that has to precede the collection of conformational constraints and the structure calculation. These methods have been reviewed recently [4], and will not be discussed in detail here. Some automated approaches [6–21] target the question of assigning the backbone and, possibly, β chemical shifts, usually on the basis of triple resonance experiments that delineate the protein backbone through one- and two-bond scalar couplings, while others [22–33] are concerned with the more demanding problem of complete assignment of the amino acid side-chain chemical shifts. In most cases, these algorithms require peak lists from a specific set of NMR spectra as input, and produce lists of chemical shifts of varying completeness and correctness, depending on the quality and information content of the input data, and on the capabilities of the algorithm.

2.2. The ambiguity of chemical shift-based NOESY assignment

In *de novo* three-dimensional structure determinations of proteins in solution by NMR spectroscopy, the key conformational data are upper distance limits derived from NOEs [34–37]. In order to extract distance constraints from a NOESY spectrum, its cross peaks have to be assigned, i.e. the pairs of interacting hydrogen atoms have to be identified. The NOESY assignment is based on previously determined chemical shift values that result from the chemical shift assignment.

Because of the limited accuracy of chemical shift values and peak positions many NOESY cross peaks cannot be attributed to a single unique spin pair but have an ambiguous NOE assignment comprising multiple spin pairs. A simple mathematical model of the NOESY assignment process by chemical shift matching gives insight into this problem [38]. It assumes a protein with n hydrogen atoms, for which complete and correct chemical shift assignments are available, and N cross peaks picked in a 2D [^{1}H,^{1}H]-NOESY spectrum with an accuracy of the peak position of $\Delta\omega$, i.e. the position of the picked peak differs from the resonance frequency of the underlying signal by no more than $\Delta\omega$ in both spectral dimensions. Under the simplifying assumption of a uniform distribution of the proton chemical shifts over a range $\Delta\Omega$, the chemical shift of a given proton falls within an interval of half-width $\Delta\omega$ about a given peak position with probability $p = 2\Delta\omega/\Delta\Omega$. Peaks with unique chemical shift-based assignment have in both spectral dimensions exactly one out of all n proton shifts inside the tolerance range $\Delta\omega$ from the peak position. Their expected number,

$$N^{(1)} = N(1-p)^{2n-2} \approx N\mathrm{e}^{-2np} = N\mathrm{e}^{-4n\Delta\omega/\Delta\Omega}, \qquad (1)$$

decreases exponentially with increasing size of the protein (n) and increasing chemical shift tolerance range ($\Delta\omega$). For a typical small protein such as the *Williopsis mrakii* killer toxin (WmKT) with

88 amino acid residues, $n = 457$ proton chemical shifts and $N = 1986$ NOESY cross peaks within a range of $\Delta\Omega = 9$ ppm [39], Eq. (1) predicts that less than 2% of the NOEs can be assigned unambiguously based solely on chemical shift information with a accuracy of $\Delta\omega = 0.02$ ppm (Fig. 1), which is an insufficient number to calculate a preliminary three-dimensional structure. For peak lists obtained from ^{13}C- or ^{15}N-resolved 3D [^{1}H,^{1}H]-NOESY spectra, the ambiguity in one of the proton dimensions can usually be resolved by reference to the hetero-spin, so that Eq. (1) is replaced by

$$N^{(1)} \approx N\mathrm{e}^{-np} = N\mathrm{e}^{-2n\Delta\omega/\Delta\Omega}. \qquad (2)$$

With regard to assignment ambiguity, ^{13}C- or ^{15}N-resolved 3D [^{1}H,^{1}H]-NOESY spectra are thus equivalent to homonuclear NOESY spectra from a protein of half the size or with twice the accuracy in the determination of the chemical shifts and peak positions.

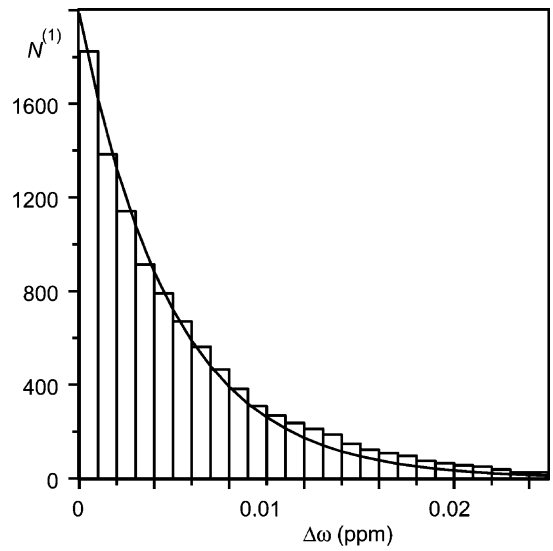


Fig. 1. Number of NOESY cross peaks with a unique chemical shift-based assignment, $N^{(1)}$, plotted as a function of the maximal chemical shift difference, $\Delta\omega$, between peak position and corresponding proton chemical shift [38]. The histogram was obtained using the experimental chemical shift list for the protein WmKT [39] and a homonuclear NOESY peak list that was simulated by postulating $N = 1986$ cross peaks for all pairs of protons that are closer than 4.0 Å in the best NMR conformer [39]. The curved line represents the corresponding values predicted by Eq. (1) for $n = 457$ proton chemical shifts, $N = 1986$ NOESY cross peaks, and $\Delta\Omega = 9.0$ ppm spectral width. No structural information was used to resolve ambiguities.

Once available, a preliminary three-dimensional structure may be used to resolve ambiguous NOE assignments. The ambiguity is resolved if only one out of all chemical shift-based assignment possibilities corresponds to an inter-atomic distance shorter than the maximal NOE-observable distance, $d_{\max}$. Assuming that the hydrogen atoms are evenly distributed within a sphere of radius R that represents the protein, the probability q that two given hydrogen atoms are closer to each other than $d_{\max}$ can be estimated by the ratio between the volumes of two spheres with radii $d_{\max}$ and R, respectively: $q = (d_{\max}/R)^3$. Using $d_{\max} = 5$ Å, one obtains $q \approx 4\%$ for WmKT, a nearly spherical protein with a radius of about 15 Å [39]. Thus, only 96% of the peaks with two assignment possibilities can be assigned uniquely by reference to the protein structure. Even by reference to a perfectly refined structure it is therefore impossible, on fundamental grounds, to resolve all assignment ambiguities, since q will always be larger than zero.

Obtaining a comprehensive set of distance constraints from a NOESY spectrum is thus by no means straightforward but becomes an iterative process in which preliminary structures, calculated from limited numbers of distance constraints, serve to reduce the ambiguity of cross peak assignments. In addition to this problem of resonance and peak overlap, considerable difficulties may arise from spectral artifacts and noise, and from the absence of expected signals because of fast relaxation. These inevitable shortcomings of NMR data collection are the main reason that until recently laborious interactive procedures have dominated 3D protein structure determinations.

2.3. Automated versus manual NOESY assignment

Automated procedures follow the same general scheme but do not require manual intervention during the assignment/structure calculation cycles (Fig. 2). Two main obstacles have to be overcome by an automated approach starting without any prior knowledge of the structure. First, the number of cross peaks with unique assignment based on chemical

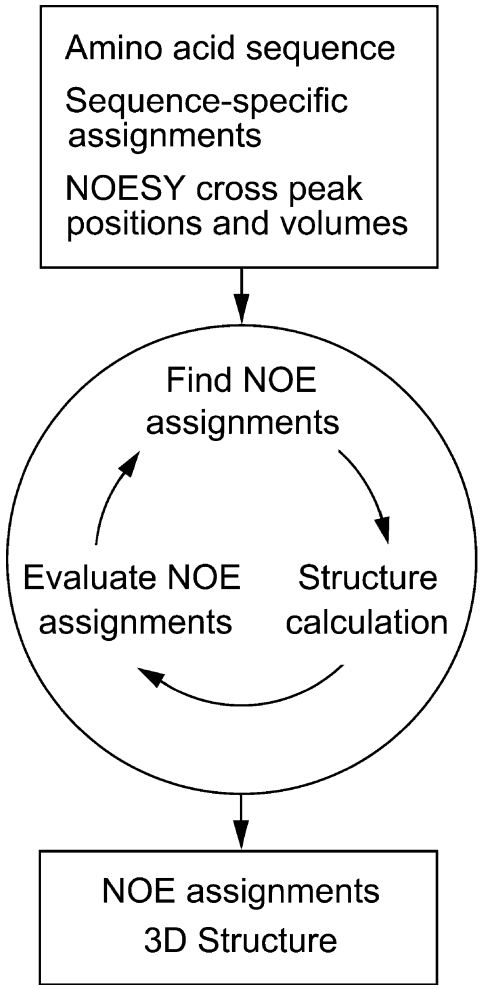


Fig. 2. General scheme of automated combined NOESY assignment and structure calculation.

shifts is, as pointed out before, in general not sufficient to define the fold of the protein. Therefore, the automated method must have the ability to make use also of NOESY cross peaks that cannot yet be assigned unambiguously. Second, the automated program must be able to cope with the erroneously picked or inaccurately positioned peaks and with the incompleteness of the chemical shift assignment of typical experimental data sets. An automated procedure needs devices to substitute the intuitive decisions made by an experienced spectroscopist in dealing with the imperfections of experimental NMR data.

3. Algorithms for automated NOESY assignment

3.1. Semi-automatic methods

Semi-automatic NOESY assignment methods relieve the spectroscopist from the burden of checking the two straightforward criteria for NOESY assignments, i.e. the agreement of chemical shifts and the compatibility with a preliminary structure, while entrusting the assignment decisions to the spectroscopist who may have additional relevant information available. Such approaches (e.g. [40–42]) use the chemical shifts and a model or preliminary structure to provide the user with a list of possible assignments for each cross peak. The user decides interactively about the assignment and/or temporary removal of individual NOESY cross peaks, possibly taking into account supplementary information such as line shapes or secondary structure data, and performs a structure calculation with the resulting, usually incomplete input. In practice, several cycles of NOESY assignment and structure calculation are required to obtain a high-quality structure.

3.1.1. The ASNO method

A prototype of this semi-automatic approach is the program ASNO [40]. The input for ASNO consists of a list of the proton chemical shifts, a peak list containing the chemical shift coordinates of the cross peaks in the NOESY spectrum, and a bundle of conformers calculated using a previous, in general preliminary set of input of NOE distance constraints. Alternatively, the structural input can consist of the crystal structure of the protein under investigation or originate from a homologous protein. However, in such applications care must be exercised to rule out possible bias by the imported reference data. In addition, the user specifies the maximally allowed chemical shift differences between corresponding cross peak coordinates and proton chemical shift values to be used for chemical shift-based assignments, the maximal proton–proton distance $d_{\max}$ in the structure that may give rise to an observable NOE, and the minimal number of conformers for which a given proton–proton distance must be shorter than $d_{\max}$ for an acceptable NOE assignment. For each NOESY cross peak ASNO first determines the set of all possible chemical shift-based assignments.

These are then checked against the corresponding $^1H-^1H$ distances in the available group of preliminary conformers and retained only if the distance between the two protons is shorter than d_{max} in at least the required number conformers. After several rounds of structure calculation, NOE assignment with ASNO, and interactive checking and refinement of the assignments, a final, high-quality structure is obtained.

3.1.2. The SANE method

The program Structure Assisted NOE Evaluation (SANE) [42] is an alternative protocol in which ambiguous distance constraints (see Section 3.3.1 below) are generated for cross peaks with multiple possible assignments. The user is directly involved in violation analysis after each round of structure calculation. Throughout the structure determination the user provides input that can help to circumvent erroneous local structures and reduce the number of iterations required to reach acceptable structures. Like ASNO, the SANE program includes a distance filter that is based on an initial search model structure, which may be an X-ray structure, an ensemble of solution structures, or even a homology-modeled structure. To minimize the problem of multiple possible assignments SANE makes use of a suite of filters that take into account existing partial assignments, the average distance between protons in one or more structures, relative NOE contributions calculated from the structures, and the expected secondary structure in order to iterate to an accurately assigned NOE cross peak list, including both unambiguous and ambiguous NOEs for the structure calculation.

3.2. The NOAH method

In a first approach and proof of feasibility of automated NOESY assignment, the programs DIANA [43] and DYANA [44] were supplemented with the automated NOESY assignment routine NOAH [38,45]. In NOAH, the multiple assignment problem is treated by temporarily ignoring cross peaks with too many (typically, more than two) assignment possibilities and instead generating independent distance constraints for each of the assignment possibilities of the remaining, low-ambiguity cross peaks, where one has to accept that part of these distance constraints may be incorrect. In order to reduce the impact of these incorrect constraints on the structure, an error-tolerant target function is used [38,45]. NOAH requires a high accuracy of the input chemical shifts and peak positions. It makes use of the fact that only a set of correct assignments can form a self-consistent network, and convergence towards the correct structure has been achieved for several proteins [38,46–48].

As an illustration, experimental 2D and 3D NOESY cross peak lists were analyzed for six proteins for which almost complete sequence-specific ^1H assignments were available for the polypeptide backbone and the amino acid side chains. The automated NOAH method assigned 70–90% of all NOESY cross peaks, which is on average 10% less than with the interactive approach, and only between 0.8 and 2.4% of the automatically assigned peaks had a different assignment than in the corresponding manually assigned peak lists. The structures obtained with NOAH/DIANA were in close agreement with those from manually assigned peak lists, and with both approaches the small remaining constraint violations indicate high-quality NMR structure determinations. Systematic comparisons of the automatically and interactively determined structures documented the absence of significant bias in either approach, indicating that an important step had been made towards automation of structure determination from NMR spectra.

In the initial assignment cycle with NOAH all peaks with one or two assignment possibilities are included into the structure calculation. In view of the large number of erroneous conformational constraints that are likely to be included at this stage, it seems non-trivial that the NOAH/DIANA approach may ultimately converge to the correct structure. The explanation is related to the fact that the structure calculation algorithm attempts to satisfy a maximum number of conformational constraints simultaneously. The correctly assigned constraints form a large subset of self-consistent constraints, whereas, in contrast, the erroneously assigned constraints are randomly distributed in space, generally contradicting each other. As a consequence, erroneously assigned constraints may distort the structure but will not lead to

a distinctly different protein fold. One must keep in mind that the elimination of erroneously assigned constraints through contradiction with correct constraints will in general be less efficient in regions of low NOE density, such as chain ends, surface loops or the periphery of long side chains, than in the well defined protein core. Another peculiarity of the randomly distributed erroneously assigned constraints is that they are more likely to be long-range than short-range or intra-residual. This contrasts with the overall constraint distribution of a correctly assigned NOESY spectrum, where more than 50% of all cross peaks are from short-range NOEs [5].

3.3. The ARIA method

The widely used automated NOESY assignment procedure ARIA [49–52] has been interfaced initially with the structure calculation program XPLOR [53] and later with the program CNS [54]. ARIA introduced many new concepts, most importantly the use of ambiguous distance constraints [55,56] for handling ambiguities in the initial, chemical shift-based NOESY cross peak assignments. Prior to the introduction of ambiguous distance constraints, in general only unambiguously assigned NOEs could be used as distance constraints in a structure calculation. Since the majority of NOEs cannot be assigned unambiguously from chemical shift information alone, this lack of a general way to directly include ambiguous data into the structure calculation considerably hampered the performance of automatic NOESY assignment algorithms.

3.3.1. Ambiguous distance constraints

When using ambiguous distance constraints, each NOESY cross peak is treated as the superposition of the signals from each of its multiple assignments, using relative weights proportional to the inverse sixth power of the corresponding inter-atomic distance. A NOESY cross peak with a unique assignment possibility gives rise to an upper bound b on the distance between two hydrogen atoms, α and β. A NOESY cross peak with $n > 1$ assignment possibilities can be seen as the superposition of n degenerate signals and interpreted as an ambiguous distance constraint, $\bar{d} \leq b$, with

$$\bar{d} = \left(\sum_{k=1}^{n} d_k^{-6} \right)^{-1/6} . \tag{3}$$

Each of the distances $d_k = d(\alpha_k, \beta_k)$ in the sum of Eq. (3) corresponds to one assignment possibility to a pair of hydrogen atoms, α_k and β_k. Because the 'r^{-6}-summed distance' $\bar{d}$ is always shorter than any of the individual distances d_k, an ambiguous distance constraint is never falsified by including incorrect assignment possibilities, as long as the correct assignment is present.

3.3.2. Overview of the ARIA algorithm

ARIA starts from lists of peaks and chemical shifts in the format of the common spectral analysis programs ANSIG [57,58], NMRView [59], PIPP [60] or XEASY [61] and proceeds in cycles of NOE assignment and structure calculation. Constraints on dihedral angles, J-couplings, residual dipolar couplings, disulfide bridges and hydrogen bonds can be used in addition, if available. In each cycle, ARIA calibrates and assigns the NOESY spectra, merges the constraint lists from different spectra, and calculates a bundle of (typically 20) conformers with the program CNS [54]. Normally, an internally generated extended start structure is used in the initial cycle 0. In all later cycles, NOE assignment, calibration and violation analysis are based on the average distances $\langle d \rangle$ calculated from the (typically 7 out of 20) lowest energy conformers from the previous cycle.

3.3.3. Calibration of distance constraints

The target distances d_{NOE} can be obtained by a simple calibration function, $d_{\mathrm{NOE}} = (CV)^{-1/6}$. The calibration constant is given by $C = \sum_{\mathrm{NOEs}} \langle d \rangle^{-6} / V$, where the sum runs over all NOEs with a corresponding average distance $\langle d \rangle$ smaller than a cutoff (typically 6 Å). An upper bound $u = d_{\mathrm{NOE}} + \varepsilon d_{\mathrm{NOE}}^2$ and a lower bound $l = d_{\mathrm{NOE}} - \varepsilon d_{\mathrm{NOE}}^2$ (typically $\varepsilon = 0.125$ Å^{-1}) are derived from each target distance d_{NOE} [51]. Alternatively, spin diffusion effects [62] can be taken into account by a relaxation matrix approach based on the simulation of the NOE spectrum rather than the direct use of the individual distances $\langle d \rangle$ [52]. A fast matrix squaring scheme performs the potentially time-consuming relaxation matrix analysis efficiently, and the deviation of the calculated NOE from the value resulting from

the isolated spin pair approximation is used to derive a correction factor for the target distance. In this way, severe cases of spin diffusion can be detected and corrected within the framework of the automated algorithm.

3.3.4. Partial NOE assignment

Despite the property of ambiguous distance constraints that additional, even wrong assignment possibilities added to an ambiguous distance constraint that contains one or several correct assignments do not render the constraint incompatible with the correct structure, it is important to reduce the ambiguity of NOE assignments as much as possible in order to obtain a well-defined structure because additional assignment possibilities 'dilute' the information contained in an ambiguous distance constraint and make it more difficult for the structure calculation algorithm to converge to the correct structure.

To this end, the relative contribution C_k of each assignment possibility to the total peak intensity is estimated from the three-dimensional structure of the previous cycle by

$$C_k = \left(\frac{\overline{\langle d \rangle}}{\langle d_k \rangle} \right)^6, \tag{4}$$

or, in the case of the relaxation matrix treatment, by the back-calculated NOE intensity [52], normalized such that the sum over all contributions to a given peak equals 1. A partial assignment is then achieved by ordering the contributions by decreasing size, and discarding the smallest contributions such that

$$\sum_{k=1}^{N_p} C_k > p, \tag{5}$$

where p is the 'assignment cutoff' and N_p the number of contributions to the peak necessary to account for a fraction of the peak volume larger than p. The parameter p is decreased from cycle to cycle and typically takes the values 1.0, 0.9999, 0.999, 0.99, 0.98, 0.96, 0.93, 0.9, 0.8 in cycles 0–8, respectively [51]. To give an intuitive meaning to the assignment cutoff p, a cross peak with two assignments may be considered [50]: If the shorter of the two distances is 2.5 Å, a value $p = 0.999$ will exclude a second distance of 7.9 Å, a value $p = 0.95$ a second distance of 4.1 Å, and a value $p = 0.8$ a second distance of 3.3 Å. If the shorter distance is 4 Å, the corresponding minimal excluded distances are 12.6, 6.6 and 5.2 Å, respectively.

3.3.5. Removal of erroneous constraints by violation analysis

Experimental peak lists can in practice not be assumed to be completely free of errors, especially in the early stages of a structure determination or if they originate from automatic peak picking. In addition, if the chemical shift assignment is incomplete, even the most carefully prepared peak list will contain peaks that cannot be assigned correctly, namely those involving unassigned spins, because the ARIA algorithm does not attempt to extend or modify chemical shift assignments provided by the user. When building a three-dimensional structure from NOE data, most erroneous distance constraints will be inconsistent with each other and with the correct ones. The erroneous constraints can therefore, in principle, be detected by analyzing the violations of constraints with respect to the bundle of three-dimensional structures from the previous cycle of calculation. The problem is to distinguish violations arising from incorrect constraints from those of correct constraints that appear as a result of insufficient convergence of the structure calculation algorithm, or as an indirect effect of structural distortions caused by other erroneous constraints. Violations due to incorrect constraints can be expected to occur in the majority of conformers rather than sporadically. Therefore, a violation analysis is performed by counting the conformers in which a given constraint is violated by more than a cutoff that is decreased gradually from 1.0 Å in the second to 0.1 Å in the final cycle of ARIA. If this is the case in more than, typically, 50% of all conformers, three options are possible [51]: The peak is either reported as a problem but still used without change, or the upper distance bound may be increased to 6 Å, or the constraint may be removed from the input for the structure calculation in the current cycle. Obviously, this kind of violation analysis can be applied only *after* a first preliminary structure has been obtained.

3.3.6. Target function with linear asymptote

In order to reduce distortions in the structures that are caused by the presence of erroneous constraints that passed undetected through this violation analysis, ARIA uses in the structure calculation with CNS a target function with a linear asymptote for large violations which limits the maximal force exerted by a violated distance constraint. The target function for a single distance constraint is [50]:

$$f(\bar{d}) = \begin{cases} (\bar{d} - l)^2 & \text{if } \bar{d} < l; \\ 0 & \text{if } l \le \bar{d} \le u; \\ (\bar{d} - u)^2 & \text{if } u < \bar{d} < u + a; \\ a(3a - 2\gamma) + \dfrac{a^2(\gamma - 2a)}{\bar{d} - u} & \\ \quad + \gamma(\bar{d} - u) & \text{if } \bar{d} \ge u + a. \end{cases} \tag{6}$$

Here, $\bar{d}$ denotes the r^{-6}-summed distance of Eq. (3), l and u are the lower and upper distance bounds, γ is the slope of the asymptotic potential, and a is the violation at which the potential switches from harmonic to asymptotic behavior.

3.3.7. Refinement in explicit solvent

Strongly simplified, 'soft' force fields are generally used for the *de novo* calculation of NMR structures. There are two reasons for this: computational efficiency and, the need to allow for a reasonably smooth folding pathway of the polypeptide chain from a random initial structure to the native conformation that is not obstructed by high energy barriers which occur if steep, divergent potentials such as the Lennard–Jones potential of standard classical molecular dynamics force fields are used. The stiffness incurred by potentials that impede the interpenetration of parts of the molecule during the initial stages of the simulated annealing procedure would result in most conformers being trapped in local minima at unfavorable energies and far from the native structure.

However, since the physical reality of the non-bonded attractive and repulsive interactions is only crudely approximated in this way, the resulting structures have often appeared to be of low quality when submitted to common structure validation programs that put much emphasis on such features as the appearance of the Ramachandran plot, staggered rotamers of side-chain torsion angles, covalent and hydrogen bond geometry, and electrostatic interactions. To remedy this situation, a short molecular dynamics trajectory in explicit solvent may be used to refine the final structure in ARIA [63]. It has been shown that a thin layer of solvent molecules around the protein is sufficient to obtain a significant improvement in validation parameters over unrefined structures, while maintaining reasonable computational efficiency [63,64].

3.3.8. Use of ARIA in practice

The ARIA algorithm is particularly efficient for improving and completing the NOESY assignment once a correct preliminary polypeptide fold is available. On the other hand, obtaining a correct initial fold at the outset of a *de novo* structure determination can be challenging because the powerful structure-based filters used for the elimination of erroneous cross peak assignments are not yet operational at that stage. It is of great help for the initial phase of the algorithm if the user can supply a limited number of already assigned long-range distance constraints. ARIA has been used in the NMR structure determinations of more than 50 proteins [51]. A similar algorithm that also relies on ambiguous distance constraints and the program XPLOR for the structure calculation has been implemented [65,66].

3.4. The AutoStructure method

An approach that uses rules for assignments similar to those used by an expert to generate an initial protein fold has been implemented in the program AutoStructure, and applied to protein structure determination [4,67]. AutoStructure is aimed at identifying iteratively self-consistent NOE contact patterns, without using any 3D structure model, and delineating secondary structures, including alignments between β-strands, based upon a combined pattern analysis of secondary structure-specific NOE contacts, chemical shifts, scalar coupling constants, and slow amide proton exchange data. The software automatically generates conformational

constraints, e.g. distance, dihedral angle and hydrogen bond constraints, and submits parallel structure calculations with the program DYANA [44]. The resulting structure is then refined automatically by iterative cycles of self-consistent assignment of NOESY cross peaks and regeneration of the protein structure with the program DYANA.

3.5. The KNOWNOE method

The program KNOWNOE [68] presents a 'knowledge-based' approach to the problem of automated assignment of NOESY spectra that is, in principle, devised to work directly with the experimental spectra without interference of an expert. Its central part is a 'knowledge-driven Bayesian algorithm' for resolving ambiguities in the NOE assignments. NOE cross peak volume probability distributions were derived for various classes of proton–proton contacts by a statistical analysis of the corresponding inter-atomic distances in 326 protein NMR structures. For a given cross peak with n possible assignments $A_1, \ldots, A_n$, the conditional probabilities $P(A_k, a|V)$ that an assignment A_k is responsible for at least a fraction a of the cross peak volume V can then be calculated from the volume probability distributions using Bayes' theorem. Peaks with one assignment A_k with a probability $P(A_k, a|V_0)$ higher than a cutoff, typically in the range 0.8–0.9, are transiently considered as unambiguously assigned. Note that a preliminary structure is not needed to achieve this discrimination, which therefore yields a higher number of unambiguous assignments than would be possible based on chemical shifts alone (see Section 2.2). With this list of unambiguously assigned peaks a set of structures is calculated. These structures are used as input for a next cycle in which only those assignments are accepted that correspond to distances shorter than a threshold $d_{\max}$, which is decreased from cycle to cycle until 5 Å, the assumed detection limit for NOEs. Since this algorithm essentially relies on the unambiguously assigned NOEs in order to calculate the intermediate structures (only for the final structure calculation are some ambiguous distance constraint used), it requires, like NOAH (see Section 3.2), a high accuracy of the chemical shifts of typically 0.01 ppm. The program KNOWNOE was tested successfully on 2D NOESY spectra of the 66 amino acid cold shock protein from *Thermotoga maritima* for which automated assignment of NOESY spectra yielded a structure of comparable quality to the one obtained from manual data evaluation [68].

3.6. The CANDID method

The CANDID algorithm [69] in the program CYANA [70] combines features from NOAH and ARIA, such as the use of three-dimensional structure-based filters and ambiguous distance constraints, with the new concepts of network-anchoring and constraint combination that further enable an efficient and reliable search for the correct fold in the initial cycle of *de novo* NMR structure determinations.

3.6.1. Overview of the CANDID algorithm

The automated CANDID method proceeds in iterative cycles of ambiguous NOE assignment followed by structure calculation with the CYANA torsion angle dynamics algorithm. Between subsequent cycles, information is transferred exclusively through the intermediary three-dimensional structures, in that the molecular structure obtained in a given cycle is used to guide the NOE assignments in the following cycle. Otherwise, the same input data are used for all cycles, that is, the amino acid sequence of the protein, one or several chemical shift lists from the sequence-specific resonance assignment, and one or several lists containing the positions and volumes of cross peaks in 2D, 3D or 4D NOESY spectra. The NOESY peak lists can be prepared either using interactive spectrum analysis programs such as XEASY [61], NMRView [59], ANSIG [57,58], or automated peak picking methods such as AUTOPSY [71] or ATNOS [72] that allow to start the NOE assignment and structure calculation process directly from the NOESY spectra. The input may further include previously assigned NOE upper distance constraints or other previously assigned conformational constraints. These will not be touched during NOE assignment with CANDID, but used for the CYANA structure calculation.

A CANDID cycle starts by generating for each NOESY cross peak an initial assignment list containing the hydrogen atom pairs that could, from the fit of chemical shifts within a user-defined tolerance range, contribute to the peak. Subsequently, for each cross peak these initial assignments are weighted with

respect to several criteria, and initial assignments with low overall score are discarded. These filtering criteria include the agreement between the values of the chemical shift list and the peak position, self-consistency within the entire NOE network (see Section 3.6.2 below), and, if available, the compatibility with the three-dimensional structure from the preceding cycle (Fig. 3). In the first cycle, network-anchoring has a dominant impact, since structure-based criteria cannot be applied yet. For each cross peak, the retained assignments are interpreted in the form of an upper distance limit derived from the cross peak volume. Thereby, a conventional distance constraint is obtained for cross peaks with a single retained assignment, and otherwise an ambiguous distance constraint is generated that embodies several assignments. Cross peaks with a poor score are temporarily discarded. In order to reduce deleterious effects on the resulting structure from erroneous distance constraints that may pass this filtering step, long-range distance constraints are incorporated into 'combined distance constraints' (see Section 3.6.3 below). The distance constraints are then included in the input for the structure calculation with the CYANA torsion angle dynamics algorithm.

The structure calculations typically comprise seven cycles. The second and subsequent cycles differ from the first cycle by the use of additional selection criteria for cross peaks and NOE assignments that are based on assessments relative to the protein 3D structure from the preceding cycle. Since the precision of the structure determination normally improves with each subsequent cycle, the criteria for accepting assignments and distance constraints are tightened in more advanced cycles of the CANDID calculation. The output from a CANDID cycle includes a listing of NOESY cross peak assignments, a list of comments about individual assignment decisions that can help to recognize potential artifacts in the input data, and a three-dimensional protein structure in the form of a bundle of conformers.

In the final CANDID cycle, an additional filtering step ensures that all NOEs have either unique assignments to a single pair of hydrogen atoms, or are eliminated from the input for the structure calculation. This allows for the direct use of the NOE assignments in subsequent refinement and analysis programs that do not handle ambiguous distance constraints.

The core of the CANDID algorithm has been implemented in the program CYANA [70]. The standard schedule and parameters for a complete automated structure determination with CYANA are specified in a script written in the interpreted command language INCLAN [44] that gives the user high flexibility in the way automated structure determination is performed without the need to modify the compiled core part of the algorithm.

3.6.2. Network-anchoring

Network-anchoring exploits the observation that the correctly assigned constraints form a self-consistent

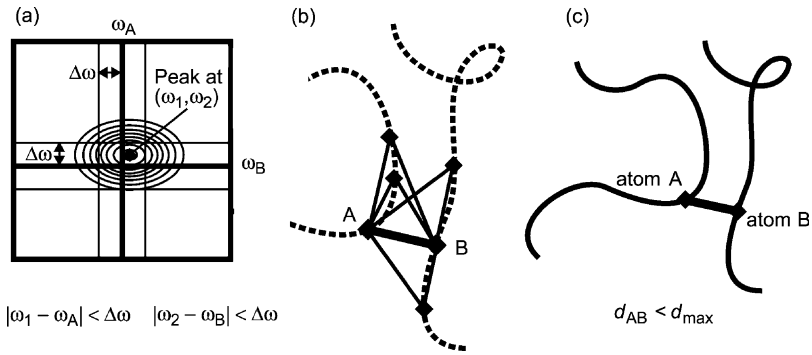


Fig. 3. Three conditions that must be fulfilled by a valid assignment of a NOESY cross peak to two protons A and B in the CANDID automated NOESY assignment algorithm [69]: (a) Agreement between chemical shifts and the peak position, (b) network-anchoring, and (c) spatial proximity in a (preliminary) structure.

subset in any network of distance constraints that is sufficiently dense for the determination of a protein 3D structure. Network-anchoring thus evaluates the self-consistency of NOE assignments independent of knowledge on the 3D protein structure, and in this way compensates for the absence of 3D structural information at the outset of a *de novo* structure determination (Fig. 3). The requirement that each NOE assignment must be embedded in the network of all other assignments makes network-anchoring a sensitive approach for detecting erroneous, 'lonely' constraints that might artificially constrain unstructured parts of the protein. Such constraints would not otherwise lead to systematic constraint violations during the structure calculation, and could therefore not be eliminated by 3D structure-based peak filters.

The network-anchoring score $N_{\alpha\beta}$ for a given initial assignment of a NOESY cross peak to an atom pair (α, β) is calculated by searching all atoms γ in the same or in the neighboring residues of either α or β that are connected simultaneously to both atoms α and β. The connection may either be an initial assignment of another peak (in the same or in another peak list) or the fact that the covalent structure implies that the corresponding distance must be short enough to give rise to an observable NOE. Each such indirect path contributes to the total network-anchoring score for the assignment (α, β) an amount given by the product of the generalized volume contributions of its two parts, $\alpha \rightarrow \gamma$ and $\gamma \rightarrow \beta$. $N_{\alpha\beta}$ has an intuitive meaning as the number of indirect connections between the atoms α and β through a third atom γ, weighted by their respective generalized volume contributions.

The calculation of the network-anchoring score is recursive in the sense that its calculation for a given peak requires the knowledge of the generalized volume contributions from other peaks, which in turn involve the corresponding network-anchored assignment contributions. Therefore, the calculation of these quantities is iterated three times, or until convergence. Note that the peaks from all peak lists contribute simultaneously to the network-anchored assignment.

3.6.3. Constraint combination

In the practice of NMR structure determination with biological macromolecules, spurious distance constraints may arise from misinterpretation of noise and spectral artifacts. This situation is particularly critical at the outset of a structure determination, before the availability of a preliminary structure for 3D structure-based filtering of constraint assignments. Constraint combination aims at minimizing the impact of such imperfections on the resulting structure at the expense of a temporary loss of information. Constraint combination is applied in the first two CANDID cycles. It consists of generating distance constraints with combined assignments from different, in general unrelated, cross peaks (Fig. 4). The basic property of ambiguous distance constraints that the constraint will be fulfilled by the correct structure whenever at least one of its assignments is correct, regardless of the presence of additional, erroneous assignments, then implies that such combined constraints have a lower probability of being erroneous than the corresponding original constraints, provided that the fraction of erroneous original constraints is smaller than 50%.

CANDID provides two modes of constraint combination (further combination modes can be envisaged readily) [69]: '$2 \rightarrow 1$' combination of all assignments of two long-range peaks each into a single constraint, and '$4 \rightarrow 4$' pairwise combination of the assignments of four long-range peaks into four constraints. Let A, B, C, D denote the sets of assignments of four peaks. Then, $2 \rightarrow 1$ combination replaces two constraints with assignment sets A and B, respectively, by a single ambiguous constraint with assignment set $A \cup B$, the union of sets A and B. $4 \rightarrow 4$ pairwise combination replaces four constraints with assignments A, B, C and D by four combined ambiguous constraints with assignment sets $A \cup B$, $A \cup C$, $A \cup D$ and $B \cup C$, respectively. In both cases constraint combination is applied only to the long-range peaks, i.e. the peaks with all assignments to pairs of atoms separated by at least five residues in the sequence, because in case of error their effect on the global fold of a protein is more pronounced than that of erroneous short-and medium-range constraints. The number of long-range constraints is halved by $2 \rightarrow 1$ combination but stays constant upon $4 \rightarrow 4$ pairwise combination. The latter approach therefore preserves more of the original structural information, and can furthermore take into account that certain peaks and their assignments are more reliable than others, because the peaks with assign-

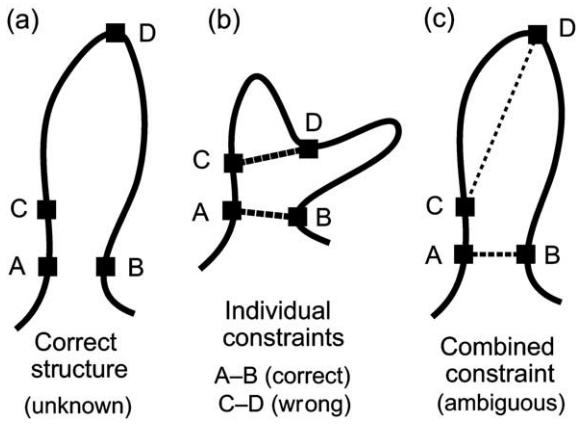


Fig. 4. Schematic illustration of the effect of constraint combination [69] in the case of two distance constraints, a correct one connecting atoms A and B, and a wrong one between atoms C and D. A structure calculation that uses these two constraints as individual constraints that have to be satisfied simultaneously will, instead of finding the correct structure (a), result in a distorted conformation (b), whereas a combined constraint that will be fulfilled already if one of the two distances is sufficiently short leads to an almost undistorted solution (c).

ment sets A, B, C, D are used 3, 2, 2, 1 times, respectively, to form combined constraints. To this end, the long-range peaks are sorted according to their total residue-wise network-anchoring and $4 \rightarrow 4$ combination is performed by selecting the assignments A, B, C, D from the first, second, third, and fourth quarter of the sorted list.

The effect of constraint combination on the expected number of erroneous distance constraints in the case of $2 \rightarrow 1$ combination may be estimated quantitatively by assuming an original data set containing N long-range peaks, and a uniform probability $p \ll 1$ that a long-range peak would lead to an erroneous constraint. By $2 \rightarrow 1$ constraint combination, these are replaced by $N/2$ constraints that are erroneous with probability p^2. In the case of $4 \rightarrow 4$ combination, it is assumed that the same N long-range peaks can be classified according to the 'safety' of their assignments into four equally large classes with probabilities αp, p, p, $(2 - \alpha)p$, respectively, that they would lead to erroneous constraints. The overall probability for an input constraint to be erroneous is again p. The parameter α, $0 \leq \alpha \leq 1$, expresses how much 'safer' the peaks in the first class are compared to those in the two middle classes, and in the fourth, 'unsafe' class. After $4 \rightarrow 4$ combination, there are still N long-range constraints but with an overall error probability of $(\alpha + (1 - \alpha^2)/4)p^2$, which is smaller than the probability p^2 obtained by simple $2 \rightarrow 1$ combination provided that the classification into more and less safe classes was successful ($\alpha < 1$). For instance, $4 \rightarrow 4$ combination will transform an input data set of 900 correct and 100 erroneous long-range cross peaks (i.e. $N = 1000$, $p = 0.1$) that can be split into four classes with $\alpha = 0.5$ into a new set of approximately 993 correct and 7 erroneous combined constraints. Alternatively, $2 \rightarrow 1$ combination will yield under these conditions approximately 495 correct and 5 erroneous combined constraints. Unless the number of erroneous constraints is high, $4 \rightarrow 4$ combination is thus preferable over $2 \rightarrow 1$ combination in the first two CANDID cycles.

The upper distance bound b for a combined constraint is formed from the two upper distance bounds b_1 and b_2 of the original constraints either as the r^{-6}-sum, $b = (b_1^{-6} + b_2^{-6})^{-1/6}$, or as the maximum, $b = \max(b_1, b_2)$. The first choice minimizes the loss of information if two already correct constraints are combined, whereas the second choice avoids the introduction of too small an upper bound if a correct and an erroneous constraint are combined.

3.6.4. Use of CANDID in practice

If used sensibly, automated NOESY assignment with CANDID has no disadvantage compared to the conventional, interactive approach but is a lot faster, and more objective. Network-anchored assignment and constraint combination render the automated CANDID method stable also in the presence of the imperfections typical for experimental NMR data sets. With CANDID, the evaluation of NOESY spectra is no longer the time-limiting step in protein structure determination by NMR. Furthermore, simple criteria based on the output of CANDID that will be given in Section 4.3 allows the reliability of the resulting structure to be assessed without cumbersome recourse to independent interactive verification of the

NOESY assignments. The CANDID method has been evaluated in test calculations [69] and used in various *de novo* structure determinations, including, for instance, four variants of the human prion protein [73,74], the pheromone binding protein from *Bombyx mori* [75], the calreticulin P-domain [76], the class I human ubiquitin-conjugating enzyme 2b [77], the heme chaperone CcmE [78] (Fig. 5), and the nucleotide-binding domain of Na, K-ATPase [79]. These structure determinations have confirmed that network-anchored assignment and constraint combination enable reliable, truly automated NOESY assignment and structure calculation without prior knowledge about NOESY assignments or the three-dimensional structure. All NOESY assignments and the corresponding distance constraints for these *de novo* structure determinations were made with CANDID, confining interactive work to the stage of the preparation of the input chemical shift and peak lists.

4. Robustness and quality control of automated NMR structure calculation

4.1. Effect of incomplete chemical shift assignments

A limiting factor for the application of all automated NOE assignment methods described in Section 3 is that they rely on the availability of an essentially complete list of chemical shifts from the preceding sequence-specific resonance assignment. At present, chemical shift assignment remains largely the domain of interactive or semi-automated methods, despite promising attempts towards automation (Section 2.1). Experience shows that in general the majority of the chemical shifts can be assigned readily whereas others pose difficulties that may require a disproportionate amount of the spectroscopist's time. Hence, NMR structure determination would be speeded up significantly if NOE assignment and structure calculation could be based on incomplete lists of assigned chemical shifts, provided that

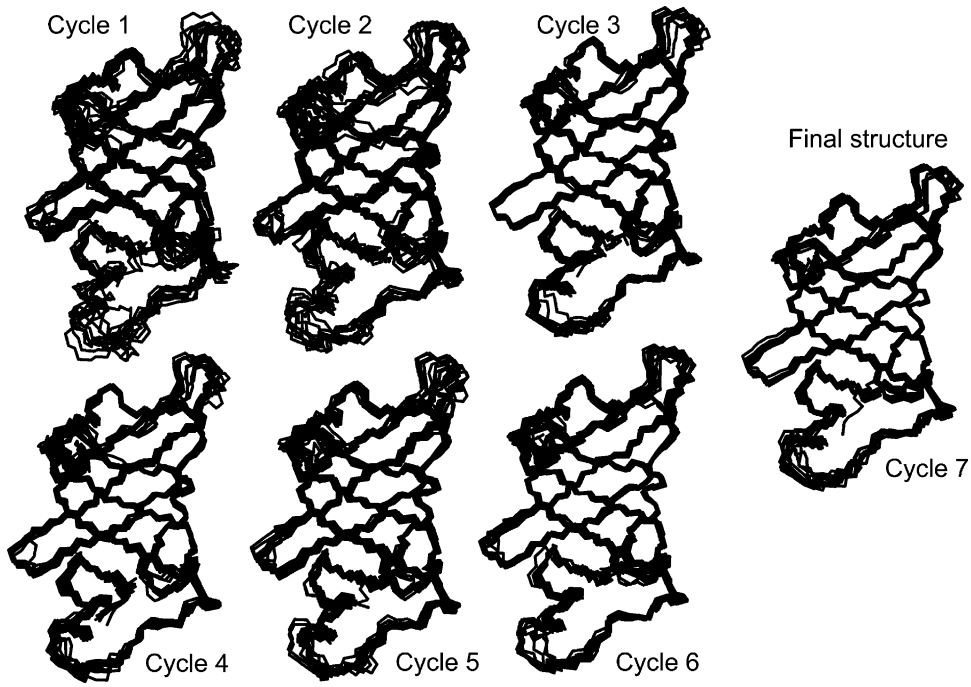


Fig. 5. Structures of the heme chaperone CcmE [78] obtained with the program CYANA [70] in seven consecutive cycles of combined automated NOESY assignment with CANDID [69] and structure calculation with torsion angle dynamics. The backbones of the 10 conformers with lowest target function value in each cycle were drawn with the program MOLMOL [94].

the reliability and robustness of the NMR method for protein structure determination is not compromised.

Methods to find additional chemical shift assignments simultaneously with automated NOESY assignment and the structure calculation have been proposed and applied with some success in the case when a preliminary structure was available [80]. For example, starting from nearly complete chemical shift assignments for the backbone and for 348 side-chain protons of the 28 kDa single-chain T cell receptor protein, the chemical shifts of 40 additional side-chain protons were found by a combination of chemical shift prediction with the program SHIFTS [81,82] and NOE assignment with ARIA [80].

The influence of incomplete chemical shift assignments on the reliability of NMR structures obtained by automated NOESY cross peak assignment has been investigated in detail [83] using the program CYANA for combined automated NOESY assignment with the CANDID algorithm and torsion angle dynamics-based structure calculations at various degrees of completeness of the chemical shift assignment. The effect of missing chemical shift assignments was simulated by randomly omitting entries from the experimental ^{1}H chemical shift lists that had been used for the earlier, conventional structure determinations of two proteins, the *Bombyx mori* pheromone binding protein form A (BmPBP^{A}) [75] and the *Williopsis mrakii* killer toxin (WmKT) [39]. Sets of structure calculations were performed with different numbers and selections of randomly omitted chemical shifts and the results compared to those obtained when using the complete experimental chemical shift list. The deviation of the structures obtained with incomplete chemical shift assignments from the reference structure was monitored by the 'RMSD bias', the RMSD between the mean coordinates of the two structure bundles [84].

In the representative case of randomly selecting the omitted chemical shifts among all ^{1}H chemical shift assignments, the RMSD bias increased only slowly with increasing omission ratio P up to about $P = 10\%$, from where onwards the RMSD bias rose abruptly, reflecting that severely distorted structures had been obtained. Higher omission ratios did not only result in high mean values of the RMSD bias but also in pronounced variations among the individual runs at a given P value with different random selections of the omitted shifts. The CYANA target function values of the final structures were, regardless of the omission ratio, almost always in the range below 5 Å^2 that is indicative of a structure that essentially fulfills all the input conformational constraints. The percentages of unassigned NOEs increased and the number of distance constraints for the final cycle of structure calculation decreased almost linearly with the omission rate. The algorithm was more tolerant against the presence of incomplete chemical shifts when run with the data from the uniformly ^{13}C- and ^{15}N-labeled protein BmPBP^{A} than with the homonuclear data for the protein WmKT despite the fact that BmPBP^{A} (142 residues) is much larger than WmKT (88 residues). This is due to the availability of ^{13}C and ^{15}N chemical shifts that allow many ^{1}H chemical shift degeneracies to be resolved, such that the probability of accidental erroneous NOE assignments is decreased compared to the case of homonuclear data. The omission of aromatic ^{1}H chemical shift assignments in general causes more severe problems than the omission of the same number of chemical shifts chosen randomly among all assigned ^{1}H chemical shifts [83]. In the case of BmPBP^{A} the omission of all assigned aromatic chemical shifts, corresponding to 6.0% of all assigned protons, led already to 2 Å RMSD bias. In the case of WmKT, with only homonuclear data, significant deviations from the reference structure were in some cases already observed at 20% omission of the aromatic chemical shifts, which corresponds to an overall omission ratio of merely 1.6% of all assigned ^{1}H chemical shifts.

Overall, the test calculations [83] show that for reliable automated NOESY assignment with the CANDID algorithm, and *a fortiori* other NOE assignment algorithms based on the same principles, around 90% completeness of the chemical shift assignment is necessary. In certain cases the lack of a small number of 'essential' chemical shifts can lead to a significant deviation of the structure. However, in practice the algorithm might be expected to tolerate a slightly higher degree of incompleteness in the chemical shift assignments than the simulations [83] suggest provided that most missing assignments are of 'unimportant' chemical shifts that are involved in only a few NOEs. This is usually the case because the chemical shifts of protons that are

involved in many NOEs, and, if absent, prevent the program from correctly assigning any of these NOEs, are intrinsically easier to assign than those exhibiting only a small number of NOEs. This effect is confirmed by the finding that the lack of aromatic chemical shifts is in general more harmful to the outcome of a structure calculation than that of a similar number of other protons because aromatic protons tend to be located in the hydrophobic core of the protein where they give rise to a higher-than-average number of NOEs.

The CANDID algorithm includes network-anchoring and constraint combination, two exclusive features that have been designed and shown to be effective in minimizing the impact of incomplete and/or erroneous pieces of input data (see Sections 3.6.2 and 3.6.3). Chemical shift assignment-based automated NOE assignment without network-anchoring and constraint combination must be expected to be more susceptible to deleterious effects from missing chemical shift assignments and artifacts in the input data.

4.2. Effect of incomplete NOESY peak picking

In contrast to the effects seen under the omission of chemical shift assignments, the random omission of NOESY peaks does not cause severe problems (Fig. 3 of Ref. [83]). Even when 50% of the NOESY peaks were omitted from the experimental input peak lists for $BmPBP^A$, most RMSD bias values remained in the region of 2 Å. An outlier with RMSD bias close to 4 Å shows that for $BmPBP^A$ the algorithm starts to loose its stability at 50% NOE omission ratio. The results with the homonuclear data from WmKT showed similar patterns, albeit with a somewhat stronger dependence on the omission rate and RMSD bias values occasionally exceeding 2 Å in runs with 30% NOESY peak omission ratio. The CYANA structure calculation protocol is thus remarkably tolerant with respect to incomplete NOESY peak picking, and can tolerate the omission of up to 50% of the NOESY cross peaks with only a moderate decrease in the precision and accuracy of the resulting structure. This suggests that it is better to strive for correctness than for ultimate completeness of the input NOESY peak lists.

4.3. Quality control

Final structures from an automatic algorithm that have a low RMSD within the bundle of conformers but differ significantly from the 'correct' reference structure are problematic because, without a knowledge of a reference structure, they may appear at first glance as good, well-defined solutions. In a conventional structure calculation based on manual NOESY assignment, incomplete or inconsistent input data will be manifested by a large RMSD and/or target function values of the final structure bundle, which will prompt the spectroscopist to correct and/or complete the input data for a next round of structure calculation. The test calculations [83] showed that for structure calculation with automated NOE assignment neither the RMSD value of the final structure nor the final target function value are suitable indicators to discriminate between correct and biased results. Other criteria are needed to evaluate the outcome.

On the basis of the initial experience with the CANDID algorithm, guidelines for successful CANDID runs were proposed [69]. These comprise six criteria that should be met simultaneously: (1) average CYANA target function value of cycle 1 below 250 Å; (2) average final CYANA target function value below 10 $Å^2$; (3) less than 20% unassigned NOEs; (4) less than 20% discarded long-range NOEs; (5) RMSD value in cycle 1 below 3 Å; and (6) RMSD between the mean structures of the first and last cycle below 3 Å. The criterion (4) refers to the percentage of NOEs discarded by the CANDID algorithm among all NOEs with assignments exclusively between atoms separated by four or more residues along the polypeptide sequence. The criteria (3) and (4) limit the number of NOEs that are not used to generate distance constraints for the final structure calculation, and thus measure the completeness with which the picked NOE cross peaks can be explained by the resulting structure.

The validity of the original guidelines as sufficient conditions for successful CANDID runs was confirmed by the fact that all the structure calculations in the systematic study [83] with an RMSD bias to the reference structure of more than 2 Å violated one or several of the six criteria. On the other hand, the test calculations [83] revealed a certain redundancy among the six original criteria. Provided that

the input peak lists do not deliberately misinterpret the underlying NOESY spectra (to which the algorithm has no direct access), the aforementioned criteria can be replaced by only two conditions. Thus, for successful structure calculation with automated NOESY assignment by the CANDID algorithm in CYANA, less than 25% of the long-range NOEs must have been discarded by the automated NOESY assignment algorithm for the final structure calculation, and the backbone RMSD to the mean coordinates for the structure bundle of the *first* cycle must not exceed 3 Å.

The percentage of discarded long-range NOEs cannot be calculated readily outside the CYANA program, because it requires knowledge of the possible assignments also for the NOESY cross peaks that were excluded from the generation of conformational constraints. In this case, an overall percentage of unused cross peaks of less than 15% can be used as an alternative criterion that is straightforward to evaluate from the final assigned output peak lists, in which unused cross peaks remain unassigned.

The ability of the program to find a well-defined structure in the initial cycle of NOE assignment and structure calculation, as measured by the RMSD within the structure bundle in cycle 1, is another important factor that strongly influences the accuracy of the final structure, as measured by the RMSD bias. This can be understood by considering the iterative nature of the CANDID algorithm, by which each cycle except cycle 1 is dependent on the structure obtained in the preceding cycle. Using network-anchoring and constraint-combination, the algorithm tries to obtain a well-defined structure already in the first cycle. A low precision of the structure from cycle 1 may hinder convergence to a well-defined final structure, or, more dangerously, opens the possibility of a structural drift in later cycles towards a precise but incorrect final structure.

4.4. Troubleshooting

If the output of a structure calculation based on automated NOESY assignment with CANDID does not fulfill these guidelines, the structure will in many cases still be essentially correct, but should not be accepted without further validation. Within the framework of CANDID, the normal approach is to improve the quality of the input chemical shift and peak lists, and to perform another CANDID run, until the criteria are met. Usually, this can be achieved efficiently because the output from an unsuccessful CANDID run, even though the structure should not be trusted per se, clearly reveals problems in the input, e.g. peaks that cannot be assigned and might therefore be artifacts or indications of erroneous or missing sequence-specific assignments. CANDID provides informational output for each peak that greatly facilitates this task: the list of its chemical shift-based assignment possibilities, the assignment(s) finally chosen, and the reasons why an assignment is chosen or not, or why a peak is not used at all. Even when the criteria of the previous section are met already, a higher precision and local accuracy of the structure might still be achieved by further improving the input data.

In principle, a *de novo* protein structure determination requires one run of CYANA with 7 cycles of automated NOE assignment and structure calculation. This is realistic when almost complete chemical shift assignments and exhaustive high-quality NOESY peak lists are available. In practice, it is often more efficient to start a first CYANA calculation from an initial, slightly incomplete list of ‘safely identifiable’ NOESY cross peaks. The results of this first CYANA calculation can then be used as additional information to prepare an improved, more complete NOESY peak list for a second CYANA calculation. This can be done more efficiently than would be possible ab initio because only peaks and regions of the protein that gave rise to problems in the first CYANA calculation need to be checked.

5. Structure calculation without chemical shift assignment

It is almost universally assumed that a protein structure determination by NMR requires the sequence-specific resonance assignments [5]. However, the chemical shift assignment by itself has no biological relevance. It is required only as an intermediate step in the interpretation of the NMR spectra. Several attempts have been made to devise a strategy for NMR protein structure determination that

circumvents the tedious chemical shift assignment step. There is an analogy between these approaches and the direct phasing methods in X-ray crystallography [85]. Although until today no *de novo* NMR protein structure determination has been accomplished without prior chemical shift assignment, an introduction into the concept of assignment-free NMR structure calculation appears warranted because recent progress in this field may open the avenue to an alternative strategy of NMR structure determination.

The underlying idea of assignment-free NMR structure calculation methods is to exploit the fact that NOESY spectra provide distance information even in the absence of any chemical shift assignments. This proton–proton distance information can be exploited to calculate a spatial proton distribution. Since there is no association with the covalent structure at this point, the protons of the protein are treated as a gas of unconnected particles. Provided that the emerging proton distribution is sufficiently clear, a model can then be built into the proton density in a manner analogous to X-ray crystallography in which the structural model is constructed into the electron density.

5.1. Initial approaches

This general idea was first tested in 1992 by Malliavin et al. [86] with 302 NOEs between backbone amide protons of lysozyme that were simulated from the crystal structure, under the assumptions that the NOEs provide distance measurements with an accuracy of $\pm 5\%$, and that the absence of a NOE indicates that the corresponding distance exceeds 4.5 Å. For the distance geometry structure calculations it was further assumed that there is no chemical shift degeneracy, i.e. it is known unambiguously whether any two pairs of NOEs involve the same proton or not. About 100 clouds of backbone hydrogen atoms were calculated using distance geometry. Despite large structural variations reflected by RMSD values of 7–14 Å among these 'structures', some secondary structure elements could be identified. Considering that even in the presence of complete chemical shift assignments the NOEs between backbone amide protons alone are in general not sufficient to determine more than a rough global fold, the results of the simulation are encouraging. Furthermore, a simplistic algorithm could extract from the proton clouds the assignments of the backbone hydrogen atoms with less than 10% error.

The question of direct structure calculation without chemical shift assignments was again investigated in 1993 by Oshiro and Kuntz [87] in simulations with synthetic NOE data for BPTI and combining metric matrix distance geometry with graph theoretical approaches to identify secondary structure elements and, eventually, sequence-specific assignments. It was concluded that 'this approach is only useful with excellent quality stereo-resolved data'.

5.2. The ANSRS method

At that time the most thorough attempt at protein three-dimensional structure determination and sequence-specific assignment of ^{13}C and ^{15}N-separated NOE data using 'a novel real-space ab initio approach' came with Per Kraulis' ANSRS algorithm in 1994 [88]. The input data are a list of NOESY cross peaks including knowledge of the chemical shifts of the ^{13}C or ^{15}N atoms covalently bound to the protons that make the NOE (i.e. a 4D NOESY peak list), and a complete but unassigned list of the chemical shifts of all detectable $^{1}H-^{13}C$ and $^{1}H-^{15}N$ moieties. The ANSRS algorithm then proceeds in three stages. First, 3D structures of unconnected ^{1}H atoms are calculated using dynamical simulated annealing. Second, a list for each residue type of plausible ^{1}H spin combinations with probability scores is generated in a recursive combinatorial search with spatial constraints. Finally, the sequence-specific assignment and a low-resolution 3D structure are obtained by Monte Carlo simulated annealing. The algorithm was tested for two small proteins, a fragment of GAL4 with 32 residues and BPTI with 58 residues using the experimental chemical shifts and synthetic NOE constraints for all distances shorter than 4 Å in the previously known 3D structures. There were 193 $^{1}H-X$ chemical shift pairs and 753 distance constraints for GAL4, and 301 H–X chemical shift pairs and 1173 distance constraints for BPTI. NOEs were interpreted in a conservative manner by using them as upper distance bounds. The resulting average 3D real-space ^{1}H spin structures were within less than 2 Å RMSD from the previously known 3D structure, and

the ANSRS procedure was able to determine the sequence-specific assignments for more than 95% of the spins. These may in turn be used as input for a conventional structure calculation in order to obtain a high-resolution structure. Despite these encouraging figures, the ANSRS program has not become a routine tool for NMR structure determination, presumably because the requirements on the quality of the input data are still formidable from the experimental point of view, and because the algorithm has no facilities to deal with overlap among $^{1}H-X$ chemical shift pairs.

5.3. Inclusion of information from through-bond spectra

Atkinson and Saudek proposed an interesting algorithm for direct fitting of structure and chemical shift data to NMR spectra [89]. Optimization of four variables per atom, three Cartesian coordinates and the chemical shift value, directly against the NOESY spectrum, rather than peak lists, by simulated annealing was shown to succeed in finding sets of coordinates (i.e. structures) and chemical shifts that match the reference configuration, albeit only in the case of a peptide fragment with six atoms. Subsequently, the same authors realized [90] that the direct determination of protein structures by NMR without chemical shift assignment is not restricted to using only NOESY spectra, but can incorporate, in a natural way, data from the same set of heteronuclear and dipolar coupling experiments as normally used in the conventional approach. NOEs are again interpreted as distances between unassigned and unconnected atoms, while cross peaks in all other spectra are also interpreted as distances instead of being used for assignment purposes. For example, a $^{15}N-^{1}H$ HSQC peak yields a distance equal to the N–H bond length between the two corresponding atoms, the HNCA spectrum yields, for each N–H pair, four distances to the two adjacent Cα atoms. To validate this principle, synthetic data was produced for the 76 amino acid protein ubiquitin: 1647 exact distances corresponding to the expected peaks from 10 heteronuclear scalar coupling experiments, 2040 4D NOE cross peaks corresponding to the $^{1}H-^{1}H$ distances shorter than 4 Å in the crystal structure, and 92,570 lower distance bounds of 4 Å for all $^{1}H-^{1}H$ distances longer than 4 Å in the crystal structure. The structure calculations with the program XPLOR yielded solutions with RMSD values to the crystal structure below 2 Å. These structures were obtained with no prior assignment of any spectral resonance or cross peak, but every hydrogen atom in the structure is labeled by both its own chemical shift and that of the attached heavy atom.

5.4. The CLOUDS method

The most recent approach to NMR structure determination without chemical shift assignment is the CLOUDS protocol of Grishaev and Llinás [91,92]. For the first time, the feasibility of the method has been demonstrated using experimental data rather than simulated data sets. The CLOUDS method relies on precise and abundant inter-proton distance constraints calculated via a relaxation matrix analysis of sets of experimental NOESY cross peaks [93]. A gas of unassigned, unconnected hydrogen atoms is condensed into a structured proton distribution (cloud) via a molecular dynamics simulated annealing scheme in which the inter-nuclear distances and van der Waals repulsive terms are the only active constraints. Proton densities are generated by combining a large number of such clouds, each computed from a different trajectory.

After filtering by reference to the cloud closest to the mean, a minimal dispersion proton density ('family of clouds', foc) is identified that affords a quasi-continuous hydrogen-only probability distribution and conveys immediate information on the shape of the protein.

The NMR-generated foc proton density provides a template to which the molecule has to be fitted to derive the structure. The primary structure is threaded through the unassigned foc by a Bayesian approach, for which the probabilities of sequential connectivity hypotheses are inferred from likelihoods of $H^{N}-H^{N}$, $H^{N}-H^{\alpha}$, and $H^{\alpha}-H^{\alpha}$ inter-atomic distances as well as ^{1}H NMR chemical shifts, both derived from public databases. Once the polypeptide sequence is identified, directionality becomes established, and the foc N and C termini are recognized. After a similar procedure, side chain hydrogen atoms are found. The folded structure is then obtained via a molecular dynamics calculation that embeds 3D structures into mirror image-related representations of the foc and selected according to a lowest energy criterion.

The feasibility of the method was tested with experimental NMR data measured for two globular protein domains, the col 2 domain of human matrix metalloproteinase-2 and the kringle 2 domain of human plasminogen, of 60 and 83 amino acid residues, respectively, for which excellent unambiguously identified homonuclear NOESY peak lists were available from the previous, conventional structure determinations. The structures deviate by 1.0–1.4 Å RMSD for the backbone heavy atoms and 1.5–2.1 Å RMSD for all heavy atoms from the previously reported X-ray and NMR structures. These results show that assignment-free NMR structure calculation can successfully generate 3D protein structures from experimental data. Nevertheless, in the course of a *de novo* structure determination it may not be straightforward to produce a NOESY peak list of the completeness and quality used for these test calculations. In particular, it was assumed that the NOEs can be identified unambiguously, i.e. that it is known with certainty whether any two NOESY peaks involve the same proton or not.

As for all NMR spectrum analysis, resonance overlap presents a major difficulty also in applying 'no assignment' strategies. Indeed, if two resonances from nuclei that are far apart in the structure have identical chemical shifts but distinct sets of neighbors they would be represented by a single atom with one set of neighbors, leading to a gross distortion of the calculated structure. In that respect, the use of heteronuclear-edited NOESY spectra drastically reduces the likelihood of overlap. At present, a full *de novo* protein structure determination by the assignment-free approach has not been reported, and it is of great interest to see whether the assignment-free approach will be able to provide the robustness and quality of the structures obtained by the conventional method.

References

[1] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, Nucleic Acids Res. 28 (2000) 235.
[2] S.E. Brenner, Nat. Rev. 2 (2001) 801.
[3] D. Vitkup, E. Melamud, J. Moult, C. Sander, Nat. Struct. Biol. 8 (2001) 559.
[4] H.N.B. Moseley, G.T. Montelione, Curr. Opin. Struct. Biol. 9 (1999) 635.
[5] K. Wüthrich, NMR of Proteins and Nucleic Acids, Wiley, 1986.
[6] M.S. Friedrichs, L. Mueller, M. Wittekind, J. Biomol. NMR 4 (1994) 703.
[7] J.B. Olson, J.L. Markley, J. Biomol. NMR 4 (1994) 385.
[8] N.E.G. Buchler, E.R.P. Zuiderweg, H. Wang, R.A. Goldstein, J. Magn. Reson. 125 (1996) 34.
[9] J.A. Lukin, A.P. Gove, S.N. Talukdar, C. Ho, J. Biomol. NMR 9 (1997) 151.
[10] D.E. Zimmerman, C.A. Kulikowski, Y. Huang, W. Feng, M. Tashiro, S. Shimotakahara, C. Chien, R. Powers, G.T. Montelione, J. Mol. Biol. 269 (1997) 592.
[11] M. Leutner, R.M. Gschwind, J. Liermann, C. Schwarz, G. Gemmecker, H. Kessler, J. Biomol. NMR 11 (1998) 31.
[12] P. Güntert, M. Salzmann, D. Braun, K. Wüthrich, J. Biomol. NMR 18 (2000) 129.
[13] H.S. Atreya, S.C. Sahu, K.V.R. Chary, G. Govil, J. Biomol. NMR 17 (2000) 125.
[14] C. Bailey-Kellog, A. Widge, J.J. Kelley, M.J. Bernardi, J.H. Bushweller, B.R. Donald, J. Comput. Biol. 7 (2000) 537.
[15] N.S. Bhavesh, S.C. Panchal, R.V. Hosur, Biochemistry 40 (2001) 14727.
[16] F. Tian, H. Valafar, J.H. Prestegard, J. Am. Chem. Soc. (2001) 11791.
[17] H.N.B. Moseley, D. Monleon, G.T. Montelione, Curr. Methods Enzymol. 339 (2001) 91.
[18] M. Andrec, R.M. Levy, J. Biomol. NMR 23 (2002) 263.
[19] A. Chatterjee, N.S. Bhavesh, S.C. Panchal, R.V. Hosur, Biochem. Biophys. Res. Commun. 293 (2002) 427.
[20] D. Monleon, K. Colson, H.N.B. Moseley, C. Anklin, R. Oswald, T. Szyperski, G.T. Montelione, J. Struct. Funct. Genom. 2 (2002) 93.
[21] B.E. Coggins, P. Zhou, J. Biomol. NMR 26 (2003) 93.
[22] C. Yu, J.F. Hwang, T.B. Chen, V.W. Soo, J. Chem. Inf. Comput. Sci. 32 (1992) 183.
[23] J. Xu, S.K. Strauss, B.C. Sanctuary, L. Trimble, J. Chem. Inf. Comput. Sci. 33 (1993) 668.
[24] J. Xu, S.K. Strauss, B.C. Sanctuary, L. Trimble, J. Magn. Reson. B103 (1994) 53.
[25] H. Oschkinat, D. Croft, Methods Enzymol. 239 (1994) 308.
[26] C. Bartels, M. Billeter, P. Güntert, K. Wüthrich, J. Biomol. NMR 7 (1996) 207.
[27] C. Bartels, P. Güntert, M. Billeter, K. Wüthrich, J. Comput. Chem. 18 (1997) 139.
[28] W.Y. Choy, B.C. Sanctuary, G. Zhu, J. Chem. Inf. Comput. Sci. 37 (1997) 1086.
[29] W. Gronwald, L. Willard, T. Jellard, R.E. Boyko, K. Rajaranthnam, D.S. Wishart, F.D. Sonnichsen, B.D. Sykes, J. Biomol. NMR 12 (1998) 395.
[30] K.B. Li, B.C. Sanctuary, J. Chem. Inf. Comput. Sci. 37 (1997) 359.
[31] K.B. Li, B.C. Sanctuary, J. Chem. Inf. Comput. Sci. 37 (1997) 467.
[32] P. Pristovšek, H. Rüterjans, R. Jerala, J. Comput. Chem. 23 (2002) 335.
[33] T.K. Hitchens, J.A. Lukin, Y. Zhan, S.A. McCullum, G.S. Rule, J. Biomol. NMR 25 (2003) 1.

[34] I. Solomon, Phys. Rev. 99 (1955) 559.
[35] S. Macura, R.R. Ernst, Mol. Phys. 41 (1980) 95.
[36] A. Kumar, R.R. Ernst, K. Wüthrich, Biochem. Biophys. Res. Commun. 95 (1980) 1.
[37] D. Neuhaus, M.P. Williamson, The Nuclear Overhauser Effect in Structural and Conformational Analysis, VCH, 1989.
[38] C. Mumenthaler, P. Güntert, W. Braun, K. Wüthrich, J. Biomol. NMR 10 (1997) 351.
[39] W. Antuch, P. Güntert, K. Wüthrich, Nat. Struct. Biol. 3 (1996) 662.
[40] P. Güntert, K.D. Berndt, K. Wüthrich, J. Biomol. NMR 3 (1993) 601.
[41] R.P. Meadows, E.T. Olejniczak, S.W. Fesik, J. Biomol. NMR 4 (1994) 79.
[42] B.M. Duggan, G.B. Legge, H.J. Dyson, P.E. Wright, J. Biomol. NMR 19 (2001) 321.
[43] P. Güntert, W. Braun, K. Wüthrich, J. Mol. Biol. 217 (1991) 517.
[44] P. Güntert, C. Mumenthaler, K. Wüthrich, J. Mol. Biol. 273 (1997) 283.
[45] C. Mumenthaler, W. Braun, J. Mol. Biol. 254 (1995) 465.
[46] Y. Xu, J. Wu, D. Gorenstein, W. Braun, J. Magn. Reson. 136 (1999) 76.
[47] Y. Xu, M.J. Jablonsky, P.L. Jackson, W. Braun, N.R. Krishna, J. Magn. Reson. 148 (2001) 35.
[48] N. Oezguen, L. Adamian, Y. Xu, K. Rajarathnam, W. Braun, J. Biomol. NMR 22 (2002) 249.
[49] M. Nilges, M.J. Macias, S.I. O'Donoghue, H. Oschkinat, J. Mol. Biol. 269 (1997) 408.
[50] M. Nilges, S.I. O'Donoghue, Prog. NMR Spectrosc. 32 (1998) 107.
[51] J.P. Linge, S.I. O'Donoghue, M. Nilges, Methods Enzymol. 339 (2001) 71.
[52] J.P. Linge, M. Habeck, W. Rieping, M. Nilges, Bioinformatics 19 (2003) 315.
[53] A.T. Brünger, X-PLOR Version 3.1. A system for X-ray crystallography and NMR, Yale University Press, 1993.
[54] A.T. Brünger, P.D. Adams, G.M. Clore, W.L. DeLano, P. Gros, R.W. Grosse-Kunstleve, J.S. Jiang, J. Kuszewski, M. Nilges, N.S. Pannu, R.J. Read, L.M. Rice, T. Simonson, G.L. Warren, Acta Crystallogr. D 54 (1998) 905.
[55] M. Nilges, Proteins 17 (1993) 297.
[56] M. Nilges, J. Mol. Biol. 245 (1995) 645.
[57] P.J. Kraulis, J. Magn. Reson. 24 (1989) 627.
[58] M. Helgstrand, P. Kraulis, P. Allard, T. Härd, J. Biomol. NMR 18 (2000) 329.
[59] B.A. Johnson, R.A. Blevins, J. Biomol. NMR 4 (1994) 603.
[60] D.S. Garrett, R. Powers, A.M. Gronenborn, G.M. Clore, J. Magn. Reson. 95 (1991) 214.
[61] C. Bartels, T. Xia, M. Billeter, P. Güntert, K. Wüthrich, J. Biomol. NMR 6 (1995) 1.
[62] A. Kalk, H.J.C. Berendsen, J. Magn. Reson. 24 (1976) 343.
[63] J.P. Linge, M.A. Williams, C.A.E.M. Spronk, A.M.J.J. Bonvin, M. Nilges, Proteins 50 (2003) 496.
[64] C.A.E.M. Spronk, J.P. Linge, C.W. Hilbers, G.W. Vuister, J. Biomol. NMR 22 (2002) 281.
[65] B. Gilquin, A. Lecoq, F. Desné, M. Guenneugues, S. Zinn-Justin, A. Ménez, Proteins 34 (1999) 520.
[66] P. Savarin, S. Zinn-Justin, B. Gilquin, J. Biomol. NMR 19 (2001) 49.
[67] N.J. Greenfield, Y.J. Huang, T. Palm, G.V.T. Swapna, D. Monleon, G.T. Montelione, S.E. Hitchcock-DeGregori, J. Mol. Biol. 312 (2001) 833.
[68] W. Gronwald, S. Moussa, R. Elsner, A. Jung, B. Ganslmeier, J. Trenner, W. Kremer, K.P. Neidig, H.R. Kalbitzer, J. Biomol. NMR 23 (2002) 271.
[69] T. Herrmann, P. Güntert, K. Wüthrich, J. Mol. Biol. 319 (2002) 209.
[70] CYANA version 1.0, www.guentert.com.
[71] R. Koradi, M. Billeter, M. Engeli, K. Wüthrich, J. Magn. Reson. 135 (1998) 288.
[72] T. Herrmann, P. Güntert, K. Wüthrich, J. Biomol. NMR 24 (2002) 171.
[73] L. Calzolai, D.A. Lysek, P. Güntert, C. von Schroetter, R. Riek, R. Zahn, K. Wüthrich, Proc. Natl Acad. Sci. USA 97 (2000) 8340.
[74] R. Zahn, P. Güntert, C. von Schroetter, K. Wüthrich, J. Mol. Biol. 326 (2003) 225.
[75] R. Horst, F. Damberger, P. Luginbühl, P. Güntert, G. Peng, L. Nikonova, W.S. Leal, K. Wüthrich, Proc. Natl Acad. Sci. USA 98 (2001) 14374.
[76] L. Ellgaard, R. Riek, T. Herrmann, P. Güntert, D. Braun, A. Helenius, K. Wüthrich, Proc. Natl Acad. Sci. USA 98 (2001) 3133.
[77] T. Miura, W. Klaus, A. Ross, P. Güntert, H. Senn, J. Biomol. NMR 22 (2002) 89.
[78] E. Enggist, L. Thöny-Meyer, P. Güntert, K. Pervushin, Structure 10 (2002) 1551.
[79] M. Hilge, G. Siegal, G.W. Vuister, P. Güntert, S.M. Gloor, J.P. Abrahams, Nat. Struct. Biol. 10 (2003) 468.
[80] B.J. Hare, G. Wagner, J. Biomol. NMR 15 (1999) 103.
[81] K. Ösapay, D.A. Case, J. Am. Chem. Soc. 113 (1991) 9436.
[82] D.F. Sitkoff, D.A. Case, J. Am. Chem. Soc. 119 (1997) 12262.
[83] J.G. Jee, P. Güntert, J. Struct. Funct. Genom. (2003) in press.
[84] P. Güntert, Q. Rev. Biophys. 31 (1998) 145.
[85] J. Drenth, Principles of Protein X-ray Crystallography, Springer, 1994.
[86] T.E. Malliavin, A. Rouh, M. Delsuc, J.-Y. Lallemand, C. R. Acad. Sci. Ser. II 315 (1992) 635.
[87] C.M. Oshiro, I.D. Kuntz, Biopolymers 33 (1993) 107.
[88] P.J. Kraulis, J. Mol. Biol. 243 (1994) 696.
[89] R.W. Atkinson, V. Saudek, J. Chem. Soc. Faraday Trans. 93 (1997) 3319.
[90] R.W. Atkinson, V. Saudek, FEBS Lett. 510 (2002) 1.
[91] A. Grishaev, M. Llinás, Proc. Natl Acad. Sci. USA 99 (2002) 6707.
[92] A. Grishaev, M. Llinás, Proc. Natl Acad. Sci. USA 99 (2002) 6713.
[93] M. Madrid, E. Llinás, M. Llinás, J. Magn. Reson. 93 (1991) 329.
[94] R. Koradi, M. Billeter, K. Wüthrich, J. Mol. Graph. 14 (1996) 51.

"Methods for Ordering Proteins"

Chung-ke Chang

Institute of Biological Sciences, Academia Sinica

The advent of residual dipolar coupling (RDC) measurements as an aid in macromolecular solution structure determination through NMR spectroscopy has sparked a tremendous interest in this technique. RDC information promises better convergence during classical structure calculations, fast fold determination suitable for high-throughput structural genomics projects using NMR techniques, and less painful determination of relative orientations between domains in large proteins. In order to achieve these, samples have to include some medium that induces a slight alignment of the protein of interest. Thus there has been an enormous amount of research done on alignment media suitable for NMR samples. Among the most well-studied and probably best known media are lipid bicelles, phage particles and polyacrylamide gels. Each medium has its own strengths and weaknesses in terms of preparation, sample suitability and spectroscopic properties. In this short survey, I will focus on the three mainstream alignment media and compare their relative advantages and disadvantages. A brief introduction will also be made on the less used media, such as cellulose-based medium and media based on organic solvents.

Partial alignment of biomolecules: an aid to NMR characterization

James H Prestegard* and Anita I Kishore

Partial alignment of biomolecules in solution has added a new dimension to structural investigation by high-resolution NMR methods. Applications to proteins, nucleic acids and carbohydrates now abound. Limitations initially associated with compatibility of biomolecules with the liquid-crystal media commonly used to achieve alignment have begun to disappear. This is, in part, a result of the introduction of a wide variety of new media. Future applications to biologically important problems such as the structural organization of multi-domain proteins and multi-protein assemblies look very promising.

Addresses
Complex Carbohydrate Research Center, University of Georgia, Athens, Georgia, 30602, USA
*e-mail: jpresteg@ccrc.uga.edu

Current Opinion in Chemical Biology 2001, 5:584–590

1367-5931/01/$ – see front matter

Abbreviations
DHPC dihexanoylphosphatidylcholine
DMPC dimyristoylphosphatidylcholine
NOE nuclear Overhauser effect

Introduction

The use of partial alignment to enhance the information available from high-resolution NMR spectra has a long history. Its roots can be traced to the substantial amount of NMR done in magnetic-field-aligned liquid crystals 30 or more years ago [1]. Thoughts about application to biomolecules in solution arose more than 15 years ago with the observation that isolated molecules with sufficiently anisotropic susceptibilities would adopt slightly non-isotropic orientational distributions when placed in very high magnetic fields [2]. Both areas of application rely on the fact that anisotropic contributions to resonance energies, such as the dipolar interaction between pairs of magnetically active spin ½ nuclei, do not average to zero when vectors defining these interactions sample non-isotropic orientational distributions during tumbling in solution. In the case of dipole–dipole interactions, splittings of resonances appear that bear an average $(3\cos^2\theta–1)/r^3$ relationship to the angle between the magnetic field and the internuclear vector, θ, and to the length of the internuclear vector, r. This relationship is rich in structural information with the angular dependence offering a particularly nice complement to the pure distance dependence of the nuclear Overhauser effect (NOE).

Despite the promise of new structural information, early attempts to exploit partial alignment met limitations that prevented widespread application to biomolecules until years later, when they were available with significant enrichment in rare isotopes such as ^{15}N and ^{13}C. In the liquid-crystal area, molecules were typically aligned quite strongly; couplings among the most easily observed magnetic nuclei (^1H) extended over long ranges, and even simple molecules such as benzene displayed spectra having dozens of lines. It would have been impossible to analyze spectra of larger biomolecules. In the field-aligned area, couplings were weak and most observable couplings were among pairs of protons with short inter-proton distances; this left the investigator with too few observables to define both parameters describing the nature of non-isotropic averaging and parameters describing angular or distance constraints. Interestingly, the high resolution available from multi-dimensional work on ^{15}N-labeled proteins gave a route to obtaining enough information on a single molecule to consider structure determination [3]. At the same time, a combination of new dilute aqueous liquid crystals and the low magnetic moment of ^{15}N spins gave the weak interactions required to allow a straightforward interpretation of residual dipolar couplings [4]. The result has been a stimulation of research into the development of new alignment media and an explosion of applications to proteins, nucleic acids and carbohydrates. The technical aspects of the collection and analysis of data have been reviewed recently [5••]; here we focus on new advances in applications, particularly with respect to use of new alignment media.

Molecular alignment

The induction of partial alignment is clearly essential to the measurement of residual dipolar couplings; the $(3\cos^2\theta–1)$ function mentioned above averages to zero in an isotropic environment. Although media capable of inducing partial alignment abound, a major obstacle to application has clearly been compatibility of biomolecules with these media. The ability to work under near-physiological conditions has always been a hallmark of NMR structural studies of biomolecules, and one clearly does not want to sacrifice this in seeking media that can induce sufficient alignment. The initial applications to proteins avoided this difficulty by relying on the inherent anisotropic magnetic susceptibility of the molecule of interest. Upon placement of any molecule in a magnetic field, a magnetic dipole moment will be induced that is proportional to its susceptibility, χ. These induced moments will in turn have an interaction energy with the magnetic field, W, that is orientationally dependent because of the anisotropic distribution of electrons in these molecules. Thus, W is written explicitly recognizing the tensoral nature of χ,

$$W = \frac{1}{\mu_0}\left(-\frac{1}{2}B.\chi.B\right) \quad (1)$$

Note that the size of the interaction energy (W) depends on the square of the magnetic field strength (B^2), and variations in W arise from the anisotropy in χ. Only if these

variations are large enough compared with the thermal energy (kT) does a measurable degree of orientational order result. For a weakly oriented system, the field dependence of the resulting residual dipolar contribution to resonance splittings can also be used to separate the contribution of interest from field-independent contributions to splittings such as scalar couplings.

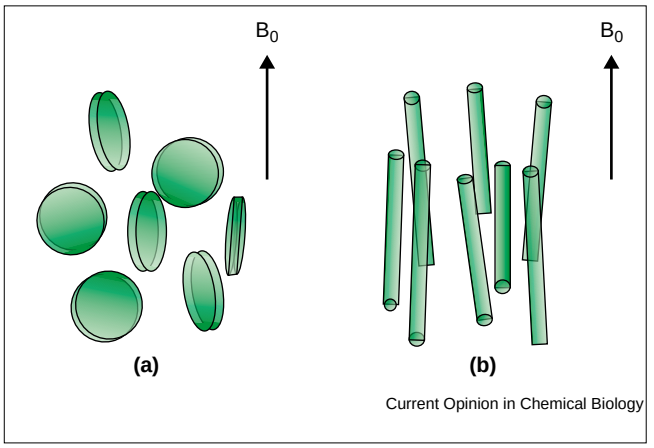
The initial application to ^{15}N-labeled proteins used a protein having a large anisotropic paramagnetic susceptibility, cyanometmyoglobin, and the highest magnetic field available at the time, 17.5 T [3]. Even so, measured ^{1}H–^{15}N splittings for directly bonded H–N pairs in amide groups only reached 4 Hz, an indication of a departure from complete isotropy of only a few parts in 10,000. Few diamagnetic assemblies (an exception is a DNA double helix) have sufficiently large anisotropies to produce splittings even this large. Fortunately, splittings of NMR resonances can be measured with great precision, and it has been possible to measure residual dipolar couplings even in some diamagnetic proteins [6]. It is important to realize, however, that when such small contributions to splittings are measured, other potential contributions to splittings, such as dynamic frequency shifts, need to be considered.

More recent applications have exploited approaches that convert normally diamagnetic proteins, or other biomolecular assemblies, to paramagnetic species. One approach uses the adventitious binding properties of lanthanides [7,8], or replaces native diamagnetic ions such as Ca^{2+} with paramagnetic ions such as Ce^{3+} [9•]. Another approach relies on the expression of chimeric constructs having small domains that contain appropriately anisotropic paramagnetic centers [10,11•]. Although these approaches may become more important as NMR spectrometers operating at higher magnetic fields become available, the relatively small levels of alignment at today's magnetic fields makes consideration of alternative methods of alignment essential. Some consideration has been given to electric field alignment, but applications are still in their infancy [12], and we confine further discussion to other alternatives.

Bicelle alignment

The most popular alternative to direct magnetic-field-induced alignment relies on molecular interactions with cooperatively ordered media, in other words, liquid crystals. If hundreds of molecules can be organized into particles in which the axes of molecular susceptibility tensors align, and if these particles are sufficiently geometrically anisotropic to allow organization into cooperative domains, nearly complete ordering at fields of typical NMR spectrometers can be achieved. The first such medium to be used with ^{15}N-labeled proteins was one consisting of lipid bilayer discs dispersed in aqueous buffers. A depiction of the medium with its preferred orientation of discs with normals perpendicular to the magnetic field is given in Figure 1. This medium is now called a bicelle medium [4]. Bicelles, which consist of a

Figure 1



Media used to align biomolecules. Preferred directions of orientation of (a) bicelles and (b) bacteriophage are shown relative to the magnetic field (arrow) [5••].

mixture of a phospholipid typical of those found in biological membranes, dimyristoylphosphatidylcholine (DMPC), and a more detergent-like lipid, dihexanoylphosphatidylcholine (DHPC), were initially developed for use with membrane-associated peptides and glycolipids [13]. However, the discovery that a bicelle medium could be diluted to produce the large aqueous spaces required for the free tumbling of soluble macromolecules provided one of the most commonly used media of today. This basic medium has now been the subject of systematic study [14], and a number of modifications have been devised to alter the temperature range over which it is aligned [15], alter the surface charge [16], alter the direction of alignment [17,18], and improve its resistance to hydrolysis [19]. Advantages include usefulness even at modest magnetic-field strengths, compatibility with many biomolecules (a bicelle does mimic a biological membrane), and a convenient nematic to isotropic transition (25°C for the DMPC system) that allows separation of scalar and dipolar contributions to splittings.

The mechanism by which biomolecules align when dissolved in a bicelle medium can be complex and includes steric interactions, electrostatic interactions, and specific surface associations. However, when steric interactions dominate, even rather simple models for collisionally induced orientation produce good agreement with observation [20•]. With bicelle media at 5% by weight in an aqueous buffer and a protein having an approximate axial ratio of 1.67:1, contributions to ^{1}H–^{15}N splittings of 20 Hz are predicted and can be verified by experiment. In cases where one is confident that steric alignment dominates, the dependence of induced orientation on molecular shape can actually provide additional structural information. More specific interactions can, however, occur and can become a disadvantage for the bicelle system. Order can become high enough to increase line widths and introduce

Table 1

Media used to align biomolecules and measure residual dipolar couplings.

Medium	Orientation, shape	Temp. range (°C)	Applications (to date)	Features and limitations	Reference
DMPC:DHPC	Perpendicular, disc	27–45	Proteins, nucleic acids, carbohydrates	Other lipids can be substituted	[4,14]
DMPC:DHPC:CTAB	Perpendicular, disc	27–40	Proteins	Positive charge	[16]
DLPC:CHAPSO	Perpendicular, disc	7–50	Proteins	Lower temperature	[15]
DIODPC:CHAPSO DIODPC:DIHPC	Perpendicular, disc	10–55	Proteins	Hydrolysis resistant due to ether linkages	[19,55]
DMPC:DHPC:DMPX	Perpendicular, disc	35–40	Membrane peptide	Negative charge	[56]
DMPC:CHAPSO	Perpendicular, disc	30–40	Proteins, glycolipids	Zwitterions	[57]
DMPC:DHPC + Ln^{3+}	Parallel, disc	35–90	Proteins	Changes director orientation	[17,58]
DBPC:DHPC	Parallel, disc	8–40	Carbohydrates	Biphenyl group	[59]
Rod-shaped viruses	Parallel, rod-like	5–60	Proteins, nucleic acids, carbohydrates	Wide temperature and concentration range	[25–27]
Cellulose crystallites	Perpendicular	37 range?	Proteins	Stable, inert	[60]
Cetylpyridinium halide/ *n*-hexanol/sodium halide	Parallel, lamellar	0–70	Proteins, carbohydrates	Sensitive to salt, Helfrich phase	[61,62]
Strained polyacrylamide gel	Mechanical orientation, gel	5–45	Proteins	Easy sample recovery	[63,64]
n-Alkyl-poly(ethylene glycol)/*n*-alkyl alcohol or glucopone/*n*-hexanol	Perpendicular, lamellar	0–40	Proteins, nucleic acids	Stable, inert	[65]
Purple membranes, rhodopsin membranes	Parallel, disc-like	< 70	Proteins, peptides	Low concentration	[24•,66,67]
Vanadium pentoxide	Parallel, ribbons	20 range?	Carbohydrates	pH < 3 negative charge	[68]

concern about selectively detecting properties of minor bicelle-associating conformers that are in rapid exchange with dominant native conformers. In most cases, it is best to avoid strong associations; this can sometimes be accomplished by adding amphiphiles to bicelles to give them a like charge to the molecule under observation [16].

There are a few occasions when strong association is desired; one example is where selective observation of properties of a small percentage of bound conformer is desired. This bears analogy to the use of transferred NOEs in NMR investigations of bound forms of protein ligands for drug-discovery projects. There are, as yet, few examples of this type of application. Two involve determination of bound orientations of saccharides interacting with lectin-like domains of carbohydrate-binding proteins [21,22]. Two involve the interaction of a small peptide with a protein receptor [23,24•]. The latter involves a membrane-associated G-protein-coupled receptor, rhodopsin, and is particularly intriguing both because of the interest in this class of molecule as a drug target and because of the use of native bicelle-like membrane particles containing the receptor.

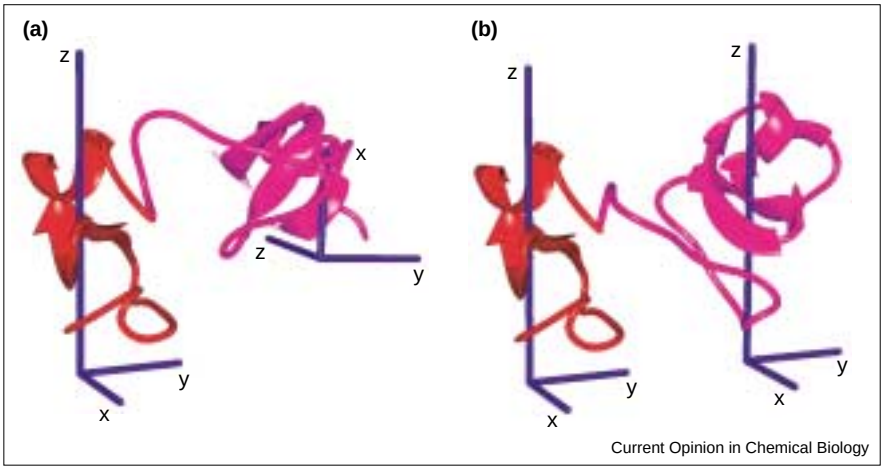
Other alignment media

In most cases, however, one still wishes to avoid association, and this has placed a premium on the discovery of other field-orientable media that have complementary surface-association properties. The first addition to widely used media came with the recognition that dispersions of rod-shaped bacteriophage could be used as a field-orientable medium [25••,26,27]. These very long rods form a cooperatively ordering medium at much lower concentrations than bicelles. They are highly negatively charged particles that induce orientation of other solutes largely through electrostatic interactions. The negative charge has made these particularly useful in the study of nucleic acids, including DNA and various types of RNA. The strong charge–charge repulsion produces order with minimal line broadening. At the same time, the strong negative charge makes this medium of little use for positively charged proteins. There have been a number of more recent additions to the arsenal of orienting media that can be used in structural characterization of biomolecules. We have attempted to summarize their properties and potential advantages in Table 1. We have taken the liberty of limiting entries to those where measurement of residual dipolar couplings on biomolecules has actually been documented.

An interesting use of multiple alignment media has occurred recently. This addresses limitations in application of residual dipolar couplings that are associated with the need to collect sufficient data to determine the five independent variables

Figure 2

Procedure for determining relative domain orientation in multi-domain proteins. (a) The alignment frame is determined from the perspective of each domain. (b) The domains are rotated so that they see a common alignment frame. The illustration uses the B and C domains from barley lectin [47].



needed to specify level of order, asymmetry of order, and the orientation of a principal order frame for a molecular fragment. Sufficient data are usually collected by measuring various types of dipolar couplings ($^{1}H–^{15}N$, $^{1}H–^{13}C$, $^{13}C–^{15}N$, etc.), but they can also be collected by measuring a more restricted set of couplings in a number of different alignment media. There are some advantages to incorporating only ^{15}N in proteins, and use of just $^{1}H–^{15}N$ and $^{1}H–^{1}H$ couplings in combination with a number of different media offers a good option when only ^{15}N enrichment is used. If enough information can be obtained on these systems, one can begin to discuss not only structure, but also internal motions (W Peti, J Meiler, R Brueschweiler, C Griesinger, unpublished data).

There are also some limitations that arise because residual dipolar couplings are insensitive to axis inversion. This usually leads to a fourfold degeneracy in one's ability to determine fragment orientation. Collection of data in multiple orienting media can remove the latter degeneracy as well [28,29]. Multiple orientating media cannot be chosen without some forethought. Using different concentrations of media or changing the media director relative to the magnetic field, for example, will not in general help; these just scale all observable couplings by a constant and provide no additional structural information. One must usually alter the order of the molecule relative to the medium particle, by changing the proportion of steric to electrostatic interactions, for example.

Applications

Along with new tools for acquiring residual dipolar coupling data have come new applications. We cannot highlight all of these, but we can mention general areas of application and emphasize one that we believe to offer much promise for the future. Refinement and determination of protein structures determined by combining residual dipolar couplings with other types of NMR data has continued [30,31]. It is now well documented that inclusion of residual dipolar coupling data increases precision, and usually accuracy, when an available X-ray structure can be used as a point of reference [5••,32•]. Rapid production of at least backbone structures of proteins based largely on residual dipolar couplings has come to the forefront because of structural genomics efforts. The fact that, unlike NOE distance constraints, the angular constraints from residual dipolar couplings do not require the close approach of constrained elements allows direct analysis of backbone geometry without complete resonance assignment for sidechains [33,34]. The ability to focus on the backbone of proteins has also been exploited in methods that search existing structural databases for shared backbone topologies [35–37]. The impact of such methods on efforts in structural genomics projects has been reviewed recently [38•].

Applications to biomolecules other than proteins have also appeared. Nucleic acids provide an interesting area of application. It has been very hard to get accurate structures for the long, extended elements in nucleic acids because of the propagation of errors when a series of short-range NOE constraints are used to determine the relative placement of the ends of helices. Recent application to such classic problems as bends in DNA duplexes [39,40], the structure of DNA quadruplexes [41], and the complete structure of a number of RNAs [42] have appeared. Applications to carbohydrates have also appeared. Like nucleic acids, NOE constraints are often sparse for these molecules and the additional structural information from residual dipolar couplings on these molecules has had a large impact [43]. One special problem that arises here, and extends to other applications, is the effect of dynamics. Carbohydrates often have substantial levels of internal motion. These are hard to quantitate using traditional spin relaxation methods because of their dependence on particular time scales of motion. Averaging of residual dipolar couplings by internal motions provides a nearly time-scale-independent measure of amplitudes of motion that can nicely complement spin-relaxation measurements. Analysis of the

effects of averaging is a complicated issue, but the first attempts at this have begun to appear for both carbohydrates [44•,45] and proteins [46•].

The one area that we would like to highlight is the use of residual dipolar coupling data in the determination of domain orientation in multi-domain proteins. This is an application that benefits in much the same way that nucleic acid structure determination benefits. Inter-domain NOEs in proteins are often sparse; although internal structures of domains are well defined by NOE constraints, positioning of domains is not. This is an area of application that seems destined to grow. It is becoming increasingly apparent that proteins do not operate in isolation. In eukaryotic organisms, multiple domain proteins abound, and many processes are mediated by multi-protein assemblies. Figure 2 illustrates one approach to using residual dipolar couplings for the determination of relative orientations of protein domains in a two-domain fragment from barley lectin [47]. Here, residual dipolar couplings of ^{15}N–^{1}H amide pairs were analyzed using the internal domain structure of a homologous protein to determine the orientation of a principal order frame as viewed from each domain (axes depicted). If the domains are part of a single rigid structure, they must be positioned so that order frame orientations are identical. In this case, some inter-domain motion was detected, so the picture represents an average inter-domain structure.

There are new approaches specifically designed to probe domain orientations and allow inclusion of other types of structural data. One uses rigid body minimization of target functions in Cartesian space [48]; another uses minimization of error functions that depend on the twist, bend, and closure angles of a hinge between domains [49]. An application to the HIV-inactivating protein cyanovirin is an interesting example of the rigid body minimization approach [50••]. It is interesting both because of its use of multiple forms of NMR data, including back-calculation of dipolar couplings based on overall molecular shape, and the fact that the resulting structures of monomer and dimer in solution are distinctly different from those seen in the X-ray structure. In fact, the solution structures can be described as ones in which like domains from the two molecules in the crystal structure of the dimer nearly switch places to form a monomer structure. Another interesting application that uses hinge minimization and employs data from multiple orientation media deals with domain orientations in T4 lysozyme [51•]. T4 lysozyme hydrolyzes a β(1–4) glycosidic linkage in a peptidoglycan substrate that binds to a cleft between two domains. Crystal structures for mutants of this enzyme show a variety of hinge angles suggesting considerable flexibility. Structure determination in solution shows a structure for the native enzyme that has a cleft which is more open than that of the corresponding crystal structure by about 17°.

The above studies suggest the potential importance of studying facile domain–domain and protein–protein interactions under solution conditions. A few years ago, the use of NMR methods to study these types of interactions would have been severely limited by an inability to work with systems of large size. Recent advances in transverse relaxation optimized spectroscopy (TROSY) have, however, pushed back molecular weight limitations for certain applications to well beyond 100 kDa [52], and TROSY methods have now been incorporated in some routines for the measurement of residual dipolar couplings [53]. Residual dipolar coupling data, and the variety of new media being developed to acquire these data, will certainly play an important role in future studies of even some multi-protein systems.

Conclusions

The measurement of anisotropic NMR parameters, such as residual dipolar couplings, through the partial alignment of biomolecules in solution has clearly come of age. Improvement of accuracy of solution structures, investigation of internal motion, and characterization of multi-component assemblies are all areas that promise future development. A key to this development will be the continued discovery of novel means of inducing low levels of alignment. We have attempted to summarize means as they exist today, but also look forward to many more contributions in the area.

Acknowledgements

This work was supported by grants from the National Institutes of Health (GM 33225) and the National Science Foundation (MCB 9726344). We thank Homayoun Valafar for his input on alignment phenomena.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Emsley JW, Lindon JC: *NMR Spectroscopy Using Liquid Crystal Solvents*. Oxford: Pergamon Press; 1975.

2. Bastiaan EW, Maclean C, Van Zilj PCM, Bothner-By AA: High-resolution NMR of liquids and gases: effects of magnetic-field-induced molecular alignment. *Annu Rep NMR Spectrosc* 1987, 19:35-77.

3. Tolman JR, Flanagan JM, Kennedy MA, Prestegard JH: Nuclear magnetic dipole interactions in field-oriented proteins — information for structure determination in solution. *Proc Natl Acad Sci USA* 1995, 92:9279-9283.

4. Tjandra N, Bax A: Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science* 1997, 278:1111-1114.

5. •• Prestegard JH, Al-Hashimi HM, Tolman JR: NMR structures of biomolecules using field oriented media and residual dipolar couplings. *Q Rev Biophys* 2001, 33:371-424.
This is a recent comprehensive review of the technical aspects of collection and analysis of residual dipolar coupling data.

6. Tjandra N, Grzesiek S, Bax A: Magnetic field dependence of N–H J splittings in ^{15}N-enriched human ubiquitin resulting from relaxation interference and residual dipolar coupling. *J Am Chem Soc* 1996, 118:6264-6272.

7. Veglia G, Opella SJ: Lanthanide ion binding to adventitious sites aligns membrane proteins in micelles for solution NMR spectroscopy. *J Am Chem Soc* 2000, 122:11733-11734.

8. Beger RD, Marathias VM, Volkman BF, Bolton PH: Determination of internuclear angles of DNA using paramagnetic assisted magnetic alignment. *J Magn Reson* 1998, 135:256-259.

9. Bertini I, Janik MBL, Liu GH, Luchinat C, Rosato A: Solution
• structure calculations through self-orientation in a magnetic field of a cerium(III) substituted calcium-binding protein. *J Magn Reson* 2001, 148:23-30.
This paper illustrates the use of field-induced orientation of a paramagnetic protein for protein structure determination.

10. Gaponenko V, Dvoretsky A, Walsby C, Hoffman BM, Rosevear PR: Calculation of z-coordinates and orientational restraints using a metal binding tag. *Biochemistry* 2000, 39:15217-15224.

11. Ma C, Opella SJ: Lanthanide ions bind specifically to an added
• 'EF-hand' and orient a membrane protein in micelles for solution NMR spectroscopy. *J Magn Reson* 2000, 146:381-384.
This paper illustrates an application using orientation of a protein engineered to have an additional paramagnetic domain.

12. Peshkovsky A, McDermott AE: Dipolar interactions in molecules aligned by strong AC electric fields. *J Magn Reson* 2000, 147:104-109.

13. Sanders CR, Hare BJ, Howard KP, Prestegard JH: Magnetically-oriented phospholipid micelles as a tool for the study of membrane-associated molecules. *Prog Nuclear Magn Reson Spectrosc* 1994, 26:421-444.

14. Ottiger M, Bax A: Characterization of magnetically oriented phospholipid micelles for measurement of dipolar couplings in macromolecules. *J Biomol NMR* 1998, 12:361-372.

15. Wang H, Eberstadt M, Olejniczak ET, Meadows RP, Fesik SW: A liquid crystalline medium for measuring residual dipolar couplings over a wide range of temperatures. *J Biomol NMR* 1998, 12:443-446.

16. Losonczi JA, Prestegard JH: Improved dilute bicelle solutions for high-resolution NMR of biological macromolecules. *J Biomol NMR* 1998, 12:447-451.

17. Prosser RS, Hunt SA, DiNatale JA, Vold RR: Magnetically aligned membrane model systems with positive order parameter: switching the sign of Szz with paramagnetic ions. *J Am Chem Soc* 1996, 118:269-270.

18. Prosser RS, Shiyanovskaya IV: Lanthanide ion assisted magnetic alignment of model membranes and macromolecules. *Concepts in Magn Reson* 2001, 13:19-31.

19. Cavagnero S, Dyson HJ, Wright PE: Improved low pH bicelle system for orienting macromolecules over a wide temperature range. *J Biomol NMR* 1999, 13:387-391.

20. Zweckstetter M, Bax A: Prediction of sterically induced alignment
• in a dilute liquid crystalline phase: aid to protein structure determination by NMR. *J Am Chem Soc* 2000, 122:3791-3792.
This paper presents a useful protocol for the prediction of alignment tensors from geometric properties of a protein and its orienting medium.

21. Bolon PJ, Al-Hashimi HM, Prestegard JH: Residual dipolar coupling derived orientational constraints on geometry in a 53kDa protein–ligand complex. *J Mol Biol* 1999, 293:107-115.

22. Shimizu H, Donohue-Rolfe A, Homans SW: Derivation of the bound-state conformation of a ligand in a weakly aligned ligand–protein complex. *J Am Chem Soc* 1999, 121:5815-5816.

23. Olejniczak ET, Meadows RP, Wang H, Cai ML, Nettesheim DG, Fesik SW: Improved NMR structures of protein/ligand complexes using residual dipolar couplings. *J Am Chem Soc* 1999, 121:9249-9250.

24. Koenig BW, Mitchell DC, Konig S, Grzesiek S, Litman BJ, Bax A:
• Measurement of dipolar couplings in a transducin peptide fragment weakly bound to oriented photo-activated rhodopsin. *J Biomol NMR* 2000, 16:121-125.
This paper describes the transfer of dipolar coupling information on a bound ligand to an exchanging pool of ligands in a manner analogous to a transferred NOE.

25. Hansen MR, Hanson P, Pardi A: Filamentous bacteriophage for
•• aligning RNA, DNA, and proteins for measurement of nuclear magnetic resonance dipolar coupling interactions. *Methods Enzymol* 2000, 317:220-240.
This is a good review of the bacteriophage alignment system and its applications to nucleic acids.

26. Hansen MR, Mueller L, Pardi A: Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions. *Nat Struct Biol* 1998, 5:1065-1074.

27. Clore GM, Starich MR, Gronenborn AM: Measurement of residual dipolar couplings of macromolecules aligned in the nematic phase of a colloidal suspension of rod-shaped viruses. *J Am Chem Soc* 1998, 120:10571-10572.

28. Ramirez BE, Bax A: Modulation of the alignment tensor of macromolecules dissolved in a dilute liquid crystalline medium. *J Am Chem Soc* 1998, 120:9106-9107.

29. Al-Hashimi HM, Valafar H, Terrell M, Zartler ER, Eidsness MK, Prestegard JH: Variation of molecular alignment as a means of resolving orientational ambiguities in protein structures from dipolar couplings. *J Magn Reson* 2000, 143:402-406.

30. Schwalbe H, Grimshaw SB, Spencer A, Buck M, Boyd J, Dobson CM, Redfield C, Smith LJ: A refined solution structure of hen lysozyme determined using residual dipolar coupling data. *Protein Sci* 2001, 10:677-688.

31. Ramirez BE, Voloshin ON, Camerini-Otero RD, Bax A: Solution structure of DinI provides insight into its mode of RecA inactivation. *Protein Sci* 2000, 9:2161-2169.

32. Clore GM, Starich MR, Bewley CA, Cai ML, Kuszewski J: Impact of
• residual dipolar couplings on the accuracy of NMR structures determined from a minimal number of NOE restraints. *J Am Chem Soc* 1999, 121:6513-6514.
This paper presents a thorough analysis of the value of residual dipolar couplings in improving the quality of protein structures.

33. Hus JC, Marion D, Blackledge M: Determination of protein backbone structure using only residual dipolar couplings. *J Am Chem Soc* 2001, 123:1541-1542.

34. Fowler CA, Tian F, Al-Hashimi HM, Prestegard JH: Rapid determination of protein folds using residual dipolar couplings. *J Mol Biol* 2000, 304:447-460.

35. Annila A, Aitio H, Thulin E, Drakenberg T: Recognition of protein folds via dipolar couplings. *J Biomol NMR* 1999, 14:223-230.

36. Delaglio F, Kontaxis G, Bax A: Protein structure determination using molecular fragment replacement and NMR dipolar couplings. *J Am Chem Soc* 2000, 122:2142-2143.

37. Andrec M, Du PC, Levy RM: Protein structural motif recognition via NMR residual dipolar couplings. *J Am Chem Soc* 2001, 123:1222-1229.

38. Prestegard JH, Valafar H, Glushka J, Tian F: Nuclear magnetic
• resonance in the era of structural genomics. *Biochemistry* 2001, 40:8677-8685.
This paper presents a discussion of the potential role of residual dipolar couplings in structural genomic projects.

39. Vermeulen A, Zhou HJ, Pardi A: Determining DNA global structure and DNA bending by application of NMR residual dipolar couplings. *J Am Chem Soc* 2000, 122:9638-9647.

40. Tjandra N, Tate S, Ono A, Kainosho M, Bax A: The NMR structure of a DNA dodecamer in an aqueous dilute liquid crystalline phase. *J Am Chem Soc* 2000, 122:6190-6200.

41. Al-Hashimi HM, Majumdar A, Gorin A, Kettani A, Skripkin E, Patel DJ: Field- and phage-induced dipolar couplings in a homodimeric DNA quadruplex, relative orientation of G•(C-A) triad and G-tetrad motifs and direct determination of C2 symmetry axis orientation. *J Am Chem Soc* 2001, 123:633-640.

42. Mollova ET, Pardi A: NMR solution structure determination of RNAs. *Curr Opin Struct Biol* 2000, 10:298-302.

43. Martin-Pastor M, Bush CA: Refined structure of a flexible heptasaccharide using ^{1}H–^{13}C and ^{1}H–^{1}H NMR residual dipolar couplings in concert with NOE and long range scalar coupling constants. *J Biomol NMR* 2001, 19:125-139.

44. Tian F, Al-Hashimi HM, Craighead JL, Prestegard JH: Conformational
• analysis of a flexible oligosaccharide using residual dipolar couplings. *J Am Chem Soc* 2001, 123:485-492.
This is one illustration of the use of residual dipolar couplings in the characterization of oligosaccharide structure.

45. Neubauer H, Meiler J, Peti W, Griesinger C: NMR structure determination of saccharose and raffinose by means of homo- and heteronuclear dipolar couplings. *Helv Chim Acta* 2001, 84:243-258.

46. • Tolman JR, Al-Hashimi HM, Kay LE, Prestegard JH: Structural and dynamic analysis of residual dipolar coupling data for proteins. *J Am Chem Soc* 2001, 123:1416-1424.
This paper discusses the impact that internal motion can have on the interpretation of residual dipolar couplings in proteins.

47. Fischer MWF, Losonczi JA, Weaver JL, Prestegard JH: Domain orientation and dynamics in multidomain proteins from residual dipolar couplings. *Biochemistry* 1999, 38:9013-9022.

48. Clore GM: Accurate and rapid docking of protein–protein complexes on the basis of intermolecular nuclear Overhauser enhancement data and dipolar couplings by rigid body minimization. *Proc Natl Acad Sci USA* 2000, 97:9021-9025.

49. Skrynnikov NR, Goto NK, Yang DW, Choy WY, Tolman JR, Mueller GA, Kay LE: Orienting domains in proteins using dipolar couplings measured by liquid-state NMR: differences in solution and crystal forms of maltodextrin binding protein loaded with beta-cyclodextrin. *J Mol Biol* 2000, 295:1265-1273.

50. •• Bewley CA, Clore GM: Determination of the relative orientation of the two halves of the domain-swapped dimer of cyanovirin-N in solution using dipolar couplings and rigid body minimization. *J Am Chem Soc* 2000, 122:6009-6016.
This is an excellent illustration of the use of residual dipolar couplings in the determination of relative domain orientation in multi-domain proteins.

51. • Goto NK, Skrynnikov NR, Dahlquist FW, Kay LE: What is the average conformation of bacteriophage T4 lysozyme in solution? A domain orientation study using dipolar couplings measured by solution NMR. *J Mol Biol* 2001, 308:745-764.
This paper presents a very thorough analysis of domain orientation and motion for a well-characterized two-domain enzyme.

52. Riek R, Pervushin K, Wuthrich K: TROSY and CRINEPT: NMR with large molecular and supramolecular structures in solution. *Trends Biochem Sci* 2000, 25:462-468.

53. Evenas J, Mittermaier A, Yang DW, Kay LE: Measurement of $^{13}C^{a}$–$^{13}C^{a}$ dipolar couplings in ^{15}N, ^{13}C, ^{2}H labeled proteins: application to domain orientation in maltose binding protein. *J Am Chem Soc* 2001, 123:2858-2864.

54. Ottiger M, Bax A: Bicelle-based liquid crystals for NMR-measurement of dipolar couplings at acidic and basic pH values. *J Biomol NMR* 1999, 13:187-191.

55. Struppe J, Whiles JA, Vold RR: Acidic phospholipid bicelles: a versatile model membrane system. *Biophys J* 2000, 78:281-289.

56. Sanders CR, Prestegard JH: Magnetically orientable phospholipid bilayers containing small amounts of a bile salt analogue, CHAPSO. *Biophys J* 1990, 78:281-289.

57. Prosser RS, Volkov VB, Shiyanovskaya IV: Solid-state NMR studies of magnetically aligned phospholipid membranes: taming lanthanides for membrane protein studies. *BioChem Cell Biol* 1998, 76:443-451.

58. Cho G, Fung BM, Reddy VB: Phospholipid bicelles with positive anisotropy of the magnetic susceptibility. *J Am Chem Soc* 2001, 123:1537-1538.

59. Fleming K, Gray D, Prasannan S, Matthews S: Cellulose crystallites: a new and robust liquid crystalline medium for the measurement of residual dipolar couplings. *J Am Chem Soc* 2000, 122:5224-5225.

60. Prosser RS, Losonczi JA, Shiyanovskaya IV: Use of a novel aqueous liquid crystalline medium for high-resolution NMR of macromolecules in solution. *J Am Chem Soc* 1998, 120:11010-11011.

61. Barrientos LG, Dolan C, Gronenborn AM: Characterization of surfactant liquid crystal phases suitable for molecular alignment and measurement of dipolar couplings. *J Biomol NMR* 2000, 16:329-337.

62. Tycko R, Blanco FJ, Ishii Y: Alignment of biopolymers in strained gels: a new way to create detectable dipole–dipole couplings in high-resolution biomolecular NMR. *J Am Chem Soc* 2000, 122:9340-9341.

63. Sass HJ, Musco G, Stahl SJ, Wingfield PT, Grzesiek S: Solution NMR of proteins within polyacrylamide gels: diffusional properties and residual alignment by mechanical stress or embedding of oriented purple membranes. *J Biomol NMR* 2000, 18:303-309.

64. Ruckert M, Otting G: Alignment of biological macromolecules in novel nonionic liquid crystalline media for NMR experiments. *J Am Chem Soc* 2000, 122:7793-7797.

65. Koenig BW, Hu JS, Ottiger M, Bose S, Hendler RW, Bax A: NMR measurement of dipolar couplings in proteins aligned by transient binding to purple membrane fragments. *J Am Chem Soc* 1999, 121:1385-1386.

66. Sass J, Cordier F, Hoffmann A, Rogowski M, Cousin A, Omichinski JG, Lowen H, Grzesiek S: Purple membrane induced alignment of biological macromolecules in the magnetic field. *J Am Chem Soc* 1999, 121:2047-2055.

67. Desvaux H, Gabriel JP, Berthault P, Camerel F: First use of a mineral liquid crystal for measurement of residual dipolar couplings of a nonlabeled biomolecule. *Angew Chem Int Ed Engl* 2001, 40:373-376.

Stereo-Array-Isotope-Labeling (SAIL) Method

-High-throughput and Accurate Structural Determinations of Proteins

Masatsune Kainosho

CREST of JST and Graduate School of Science, Tokyo Metropolitan University

Conventional protein structural determinations by NMR utilize uniformly ^{13}C and/or ^{15}N labeled samples. However, the molecular size limits of the proteins to be studied remain ~ 30 kDa and structural determinations of proteins around this size-limit still remain quite challenging, if possible. This is where deuterium labeling can play a major role again. Methods such as random fractional deuteration or selective protonation, which have already been tested, are compromises and extend the molecular size limit at the expense of signal sensitivity and the accuracy of the resultant structure. More robust and uncompromised techniques should therefore be exploited, if NMR spectroscopy is to remain to be a competitive method for structural determinations of larger proteins and protein complexes. During the 40 year history of biological NMR spectroscopy, it has been clear that concomitant advances in spectroscopic methods and in preparative methods of isotopically labeled proteins are essential to overcome the numerous difficulties. Here I propose an innovative new strategy named stereo-array-isotope-labeling (SAIL) and present some of the recent results obtained using this strategy.

Residual Dipolar Couplings in the Selection and Characterization of Structural Genomics Targets

J. H. Prestegard

Southeast Collaboratory for Structural Genomics, University of Georgia, USA

One of the underlying tenants of the structural genomics program in the US is that producing representative protein structures in each "fold family" will allow computational prediction of structures for a large percentage of new genomic products. This places a premium on both the early identification of proteins representing new folds and on the efficient production of quality structures at the backbone level. Since computational modeling is done with a small percentage of sequence identity (30%), it is the backbone structures that prove most useful. Residual dipolar couplings offer significant potential in these identification and structure production activities. Couplings from backbone sites, such as one bond ^{15}N-^{1}H couplings, are easily acquired and used in pattern matching algorithms to identify proteins that do not belong to existing fold families. ^{15}N-^{1}H couplings can also be combined with other backbone couplings to provide important angular constraints on the geometry adopted by peptide units in novel proteins. In favorable cases complete backbone structures can be derived with little additional information. Progress on structure determination of target proteins selected by the Southeast Collaboratory for Structural Genomics will be discussed.

November 2nd

Handout for laboratory

(B1A Conference Room)

TATAPRO

Version 1

H.S.Atreya & K.V.R.Chary

Department of Chemical Sciences
Tata Institute of Fundamental Research
Homi Bhabha Road, Colaba, Mumbai, 400 005, India

Acknowledgements

We gratefully acknowledge the facilities provided by the National Facility for High Field NMR, supported by Department of Science and Technology (DST), New Delhi, India, Department of Biotechnology (DBT), New Delhi, India, Council of Scientific and Industrial Research (CSIR), New Delhi, India, and Tata Institute of Fundamental Research, Mumbai, India.

Recommended Citation

When citing TATAPRO Version 1 in the literature, the following citation should be used:

H.S. Atreya, S. C. Sahu, K. V. R. Chary and Girjesh Govil, A Tracked Approach for Automated NMR Assignments in Proteins (TATAPRO). J. Biomol. NMR. **17**, 125-136, 2000.

1. Introduction:

The algorithm, TATAPRO is a novel automated approach for the sequence specific NMR assignments of $^{1}H^{N}$, ^{15}N, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'/^{1}H^{\alpha}$ spins in proteins, using triple resonance experimental data. It utilizes the protein primary sequence and peak lists from a set of triple resonance spectra which correlate $^{1}H^{N}$ and ^{15}N chemical shifts with those of $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'/^{1}H^{\alpha}$. The information derived from the correlations mentioned above is used to create a *"master_list"* consisting of all possible sets of $^{1}H^{N}_{i}$, $^{15}N_{i}$, $^{13}C^{\alpha}_{i}$, $^{13}C^{\beta}_{i}$, $^{13}C'_{i}/^{1}H^{\alpha}_{i}$, $^{13}C^{\alpha}_{i-1}$, $^{13}C^{\beta}_{i-1}$ and $^{13}C'_{i-1}/\ ^{1}H^{\alpha}_{i-1}$ chemical shifts. On the basis of an extensive statistical analysis of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shift data of proteins derived from BioMagResBank (BMRB), the 20 amino acid residues that constitute proteins are grouped into 8 distinct categories, each of which is assigned a unique 2-digit code. Such a code is used to tag individual sets of chemical shifts in the master_list and also to translate the protein primary sequence into an array called pps_array. The program then uses the master_list to search for neighbouring partners of a given amino acid residue along the polypeptide chain and sequentially assigns maximum possible stretch of residues on either side. While doing so, each assigned residue is tracked in an array called assig_array, with the 2-digit code assigned earlier. Such tracked assig_array is then mapped onto the pps_array for sequence specific resonance assignments. The program has been tested using experimental data on a calcium binding protein from *Entamoeba histolytica* (*Eh*-CaBP, 15 kDa) having substantial internal sequence homology and using published data on four other proteins in the molecular weight range of 18-42 kDa. In all the cases, nearly complete sequence specific resonance assignments (> 95%) are obtained. Furthermore, the reliability of the program has been tested by deleting sets of chemical shifts randomly from the master_list created for the test proteins.

2. Description of the algorithm:

First, read ***"README"*** file. The algorithm can be divided into three important steps namely, ***master_list*** preparation, assignment of 2-digit codes to the rows in the ***master_list*** and to the individual amino acid residues in the primary sequence and finally, carrying out sequence specific resonance assignments. These steps are described below:

2.1. "Mater_list" preparation:

Peak lists ***cbcaconh.pl, cbcanh.pl, hncaco.pl and hnco.pl*** derived from CBCANH, CBCA(CO)NH, HN(CA)CO and HNCO(or alternatively, CBCANH, CBCA(CO)NH, HN(CA)HA and HN(COCA)HA) spectra, respectively, are used to group the chemical shifts and prepare the ***master_list*** as follows.

2.1.1. cbcaconh.pl:

This is an automatically picked CBCA(CO)NH spectral peak list that has information about $^{13}C^{\alpha}_{i-1}$ and $^{13}C^{\beta}_{i-1}$ chemical shifts for a given pair of $^{15}N_{i}$ and $^{1}H^{N}_{i}$. From such a list, the chemical shifts of $^{13}C^{\alpha}_{i-1}$ and $^{13}C^{\beta}_{i-1}$ are identified for each specific pair of $^{15}N_{i}$ and $^{1}H^{N}_{i}$ chemical shifts within the user defined tolerance limits (see input file ***pml.input***) and grouped into individual sets. These sets are listed in a frequency list named as ***cbcaconh.fl.***

2.1.2. cbcanh.pl:

This is an automatically picked CBCANH peak list that has information about $^{13}C^{\alpha}_{i}$, $^{13}C^{\beta}_{i}$, $^{13}C^{\alpha}_{i-1}$ and $^{13}C^{\beta}_{i-1}$ chemical shifts for a given pair of $^{15}N_i$ and $^{1}H^{N}_{i}$ chemical shifts. For each set of $^{15}N_i$, $^{1}H^{N}_{i}$, $^{13}C^{\alpha}_{i-1}$ and $^{13}C^{\beta}_{i-1}$ chemical shifts listed in the ***cbcaconh.fl***, the search is now carried out in the CBCANH peak list to identify $^{13}C^{\alpha}_{i}$ and $^{13}C^{\beta}_{i}$ chemical shifts, within the user defined tolerance limits (see input file ***pml.input***). This helps in grouping $^{1}H^{N}_{i}$, $^{15}N_i$, $^{13}C^{\alpha}_{i}$, $^{13}C^{\beta}_{i}$, $^{13}C^{\alpha}_{i-1}$ and $^{13}C^{\beta}_{i-1}$ chemical shifts as individual sets in a frequency list named as ***cbcanh_cbcaconh.fl***.

2.1.3. hncaco.pl and hnco.pl:

These are the automatically picked HN(CA)CO and HNCO peak lists that has information about $^{13}C'_{i}$ and $^{13}C'_{i-1}$ chemical shifts, respectively, for a given pair of $^{15}N_i$ and $^{1}H^{N}_{i}$ chemical shifts. Once $^{1}H^{N}_{i}$, $^{15}N_i$, $^{13}C^{\alpha}_{i}$, $^{13}C^{\beta}_{i}$, $^{13}C^{\alpha}_{i-1}$ and $^{13}C^{\beta}_{i-1}$ chemical shifts are grouped into individual sets in ***cbcanh_cbcaconh.fl***, $^{13}C'_{i}$ and $^{13}C'_{i-1}$ chemical shifts are obtained using ***hncaco.pl and hnco.pl***, respectively, for every pair of $^{1}H^{N}_{i}$ and $^{15}N_i$ chemical shifts, within the user defined tolerance limits (see input file ***pml.input***).

2.1.4. The "master_list":

The methodology described above results in grouping of individual chemical shifts into a peak list containing sets of $^{1}H^{N}_{i}$, $^{15}N_i$, $^{13}C^{\alpha}_{i}$, $^{13}C^{\beta}_{i}$, $^{13}C'_{i}$, $^{13}C^{\alpha}_{i-1}$, $^{13}C^{\beta}_{i-1}$, and $^{13}C'_{i-1}$ chemical shifts. This list is called as the ***"master_list".*** This forms the input for the next step in the algorithm.

2.2. Assignment of 2-digit codes:

As discussed in our paper, the amino acid residues are classified into 8 (now into 9) different categories based on their characteristic $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts. The 2-digit code assigned to individual amino acid residues (Table 1) is used to tag the

Table 1

Sr. No	*$^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts (δ in ppm)*	*Amino acid residues*	*2 digit code*
1	Absence of $^{13}C^{\beta}$	Gly	1 0
2	$15 < \delta(^{13}C^{\beta}) < 24$	Ala	2 0
3	$58 < \delta(^{13}C^{\beta}) < 67$	Ser	3 0
4	$24 < \delta(^{13}C^{\beta}) < 36$ & $\delta(^{13}C^{\alpha}) < 64$	Lys, Arg, Gln, Glu, His, Trp, Cys^{Red}, Val & Met	4 0
5	$24 < \delta(^{13}C^{\beta}) < 36$ & $\delta(^{13}C^{\alpha}) \geq 64$	Val	4 1
6	$36 < \delta(^{13}C^{\beta}) < 50$ & $\delta(^{13}C^{\alpha}) < 64$	Asp, Asn, Phe, Tyr, Cys^{Oxd}, Ile & Leu	5 0
7	$36 < \delta(^{13}C^{\beta}) < 50$ & $\delta(^{13}C^{\alpha}) \geq 64$	Ile	5 1
8	------	Pro	6 0
9	$\delta(^{13}C^{\beta}) > 67$	Thr	7 0

individual ***rows*** in the ***master_list*** depending on the observed $^{13}C^{\alpha}_{i}$ and $^{13}C^{\beta}_{i}$ chemical shift values.

Simultaneously, all the amino acid residues in the protein primary sequence are assigned the 2 digit codes given in Table 1. Thus, on assigning these codes the ***protein primary sequence*** gets translated into an array of 2-digit codes referred as ***pps_array*** (see below):

- ***Protein primary sequence***

  M A E A L F K E I D V N G D G A V S Y E E V K A F V S K K R A I K N E Q L L Q L I F K S I D A D G N G E I D Q N E F A K F Y G S I Q G Q D L S D D K I G L K V L Y K L M D V D G D G K L T K E E V T S F F K K H G I E K V A E Q V M K A D A N G D G Y I T L E E F L E F S L

- ***pps_array (Protein primary sequence array):***

  4 0 2 0 4 0 2 0 5 0 5 0 4 0 4 0 5 1 5 0 4 1 5 0 1 0 5 0 1 0 2 0 4 1 3 0 5 0 4 0 4 0 4 1 4 0 2 0 5 0 4 1 3 0 4 0 4 0 4 0 2 0 5 1 4 0 5 0 4 0 4 0 5 0 5 0 4 0 5 0 5 1 5 0 4 0 3 0 5 1 5 0 2 0 5 0 1 0 5 0 1 0 4 0 5 1 5 0 4 0 5 0 5 0 5 0 2 0 4 0 5 0 5 0 1 0 3 0 5 1 4 0 1 0 4 0 5 0 5 0 3 0 5 0 5 0 4 0 5 1 1 0 5 0 4 0 4 1 5 0 5 0 4 0 5 0 4 0 5 0 4 1 5 0 1 0 5 0 1 0 4 0 5 0 3 0 4 0 4 0 4 0 4 1 3 0 3 0 5 0 5 0 4 0 4 0 4 0 1 0 5 1 4 0 4 0 4 1 2 0 4 0 4 0 4 1 4 0 4 0 2 0 5 0 2 0 5 0 1 0 5 0 1 0 5 0 5 1 3 0 5 0 4 0 4 0 5 0 5 0 4 0 5 0 3 0 5 0

2.3. Sequence specific resonance assignments:

The algorithm uses the ***master_list*** and the user defined tolerance limits (see input file ***assign.input***) for sequence specific resonance assignments. As described earlier, each row in the master_list consists of $^{1}H^{N}_{i}$, $^{15}N_{i}$, $^{13}C^{\alpha}_{i}$, $^{13}C^{\beta}_{i}$, $^{13}C'_{i}$, $^{13}C^{\alpha}_{i-1}$, $^{13}C^{\beta}_{i-1}$, and $^{13}C'_{i-1}$ chemical shift values. To begin with, the algorithm reads in the $^{13}C^{\alpha}_{i}$, $^{13}C^{\beta}_{i}$ and $^{13}C'_{i}$ chemical shift values from the first row in the ***master_list*** and searches for a row where, within the user defined tolerance limits (see input file ***assign.input***), these three chemical shifts are seen as $^{13}C^{\alpha}_{i-1}$, $^{13}C^{\beta}_{i-1}$, and $^{13}C'_{i-1}$ chemical shifts. If the search is successful, the 2-digit code associated with the new row is stored in an ***assig_array***. This procedure corresponds to forward assignment in the primary sequence, which is continued till a break is encountered. Once a stretch of amino acid residues has been assigned in the forward direction, the algorithm continues with the assignment in the backward direction starting again from the first row in the ***master_list***. For backward assignment, the program reads in the $^{13}C^{\alpha}_{i-1}$, $^{13}C^{\beta}_{i-1}$, and $^{13}C'_{i-1}$ chemical shifts for a given row in the ***master_list*** and searches for the row, where these chemical shifts are seen as $^{13}C^{\alpha}_{i}$, $^{13}C^{\beta}_{i}$ and $^{13}C'_{i}$ chemical shifts, within the user defined tolerance limits. If the search is successful, the 2-digit code associated with the new row is stored in the same ***assig_array***, as was done in the case of forward assignment. The assignment is continued till a break is encountered. Thus, after assigning the residues in both forward and backward directions, the program maps the ***assig_array*** onto the ***pps_array***. A one-to-one correspondence with the ***pps_array*** results in the sequence specific resonance assignment of that polypeptide stretch. Following this, all the assigned rows are deleted from the ***master_list*** before the next round of assignment commences, for which the first row in the updated ***master_list*** is chosen as the next starting point.

2.4. Assignment of a lone residue flanked by two polypeptide segments:

This is the last step in the assignment procedure. During the process of sequence specific resonance assignments described above, one may end up with several unassigned lone residues other than prolines, that are flanked by assigned polypeptide stretches. In such an event, the information about the $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ chemical shifts of the residue preceding the unassigned one, is used to assign the ^{15}N and ^{1}H chemical shifts of the latter by utilizing CBCA(CO)NH and HNCO peak lists namely ***cbcaconh.pl*** and ***hnco.pl***. On the other hand, the $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ chemical shifts of all unassigned lone residues and that of Pro residues are obtained from the ***row*** corresponding to their succeeding residue in the ***master_list***.

2.5. The executable files:

pml *(prepares the master file)*
assign *(carries out the sequence specific resonance assignment)*
gaps *(fills up the gaps)*

2.6. How to use the programs ?

First, copy and edit ***pml.input*** and ***assign.input*** files as described in the example files. Also, create ***pps*** (protein primary sequence). Execution of the command ***"pml"*** automatically prepares the ***"master_list".*** Once the ***master_list*** is ready, execution of the command ***"assign"*** starts the sequence specific resonance assignment. This assignment procedure is carried out in several rounds ***(See the illustrative example given at the end of this Manual, section 2.8.12)***. In each round, only 20-30 amino acid residues are assigned. In this procedure, to start with one has to use stringent tolerance limits or to avoid multiple pathways of resonance assignment or to keep to them to a minimum, which are indeed searched and recorded during the course of assignment. Such stringent tolerance limits are retained till 40-50% of the amino acid residues are assigned. Thereafter, the tolerance limits are loosened and the next rounds of assignment is carried out. Lastly, unassigned lone residues, if any, that are flanked by assigned polypeptide stretches, are assigned by the execution of the command ***"gaps".***

2.7. The OUTPUT:

At the end of the assignment procedure described above the output comes out as a list of chemical shifts, which is named as ***"assignment.list"***. The output also contains a list of mismatches, which is named as ***"mismatch.out"***. This highlights the assigned amino acid residues which have unusual $^{13}C^{\beta}$ chemical shifts.

2.8. Illustrative examples and format of peak lists

(These are extracts of the triple resonance spectral data of a calcium binding protein from Entamoeba histalytica(15 kDa; 134 amino acids))

2.8.1. cbcaconh.pl

CBCA(CO)NH peak-list*

Expected number of peaks ~ 134 * 2 = 268.

Number of automatically picked peaks were 1144

S. No	$^{1}H^{N}$	^{15}N	^{13}C	Volume
1	4.944	128.217	82.241	184751.00
2	4.934	135.193	82.110	221148.59
3	4.933	124.677	82.032	177842.20
4	4.878	124.772	82.118	170538.39
5	4.920	136.008	81.810	162020.41
.	.	.	.	.
.	.	.	.	.
1142	7.651	124.781	13.689	207577.70
1143	5.212	119.500	13.498	156760.39
1144	4.784	115.041	13.524	158999.00

*2.8.2. cbcanh.pl**

CBCANH peak-list

Expected number of peaks ~ 134 * 4= 536

Number of automatically picked peaks were 3438

S. No.	$^{1}H^{N}$	^{15}N	^{13}C	Volume
1	7.365	122.479	83.095	-653193.88
2	7.756	124.960	82.816	-367818.81
3	10.745	124.574	82.804	334473.31
4	11.264	112.615	82.823	320483.22
5	8.446	112.456	82.815	302091.91
.	.	.	.	.
.	.	.	.	.
3437	6.661	115.901	13.529	-331251.41
3438	4.938	116.022	13.509	-395655.91

****The 3D peaks have been picked from the respective 3D triple resonance spectra at a low threshold. The acceptable threshold can be the one which will not pick more than 10,000 peaks in a CBCANH spectrum for any given protein. NOTE: One should not peak-pick in individual 2D planes of 3D spectra. Peak-picking refers to 3D-peak picking.***

*2.8.3. hnco.pl**

HNCO peak-list
Expected number of peaks ~ 134 * 1= 134
Number of automatically picked peaks were 172

S. No.	$^{1}H^{N}$	^{15}N	^{13}C	Volume
1	8.196	125.734	180.395	5355949.00
2	7.713	124.023	180.256	6045804.00
3	7.824	123.806	180.250	532672.81
4	7.508	128.178	180.130	412123.00
5	6.801	128.233	180.126	342663.28
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
171	7.718	119.249	173.027	5713476.00
172	7.757	119.079	173.044	5686792.00

*2.8.4. hncaco.pl**

HN(CA)CO peak-list
Expected number of peaks ~ 134 * 1= 134
Number of automatically picked peaks were 300

S. No.	$^{1}H^{N}$	^{15}N	^{13}C	Volume
1	7.426	133.200	181.761	5696070.00
2	10.232	117.372	181.653	417542.19
3	4.908	115.653	181.316	410516.28
4	11.111	134.013	181.088	423678.59
5	4.989	134.867	180.962	393169.69
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
299	4.921	130.418	166.175	418544.19
300	6.980	117.390	179.330	818393.00

The 3D peaks have been picked from the respective 3D triple resonance spectra at a low threshold. The acceptable threshold can be the one which will not pick more than 10,000 peaks in a CBCANH spectrum for any given protein. NOTE: One should not peak-pick in individual 2D planes of 3D spectra. Peak-picking refers to 3D-peak picking.

2.8.5. cbcaconh.fl

Individual sets of $^{13}C^{\alpha}_{i-1}$ & $^{13}C^{\beta}_{i-1}$ chemical shifts for a given pair of $^{15}N_i$ & $^{1}H^{N}_{i}$ as derived from CBCA(CO)NH peak list

S.No.	$^{15}N_i$	$^{1}H^{N}_{i}$	$^{13}C^{\alpha}_{i-1}$	$^{13}C^{\beta}_{i-1}$
1	129.50	8.21	68.80	30.65
2	118.64	8.38	68.83	40.00
3	123.55	8.70	67.71	31.29
4	123.44	8.60	67.68	31.29
.	.	.	.	.
.	.	.	.	.
321	125.34	7.75	45.20	00.00
322	121.58	7.93	45.27	00.00
323	118.87	7.77	45.25	00.00

2.8.6. cbcanh_cbcaconh.fl

Individual sets of $^{13}C^{\alpha}_{i-1}$/$^{13}C^{\alpha}_{i}$ & $^{13}C^{\beta}_{i-1}$/$^{13}C^{\beta}_{i}$ chemical shifts for a given pair of $^{15}N_i$ & $^{1}H^{N}_{i}$ as derived from CBCANH peak-list

S.No.	$^{15}N_i$	$^{1}H^{N}_{i}$	$^{13}C^{\alpha}_{i}$	$^{13}C^{\beta}_{i}$	$^{13}C^{\alpha}_{i-1}$	$^{13}C^{\beta}_{i-1}$
1	111.46	7.83	47.15	0.00	53.04	39.60
2	111.68	7.87	47.01	0.00	56.33	31.77
3	111.85	7.88	47.01	0.00	56.33	39.73
4	111.85	7.88	47.01	0.00	53.04	39.73
5	111.89	7.54	47.56	0.00	51.80	37.55
.	.	.	.	.	.	.
.	.	.	.	.	.	.
213	133.41	8.25	56.33	42.47	54.13	39.05
214	135.21	7.96	54.55	20.25	53.45	38.90
215	135.34	8.30	46.86	0.00	60.99	38.36

2.8.7. hnco_hncaco.fl

Individual sets of $^{13}C'_{i-1}$ & $^{13}C'_{i}$ chemical shifts for a given pair of $^{15}N_i$ & $^{1}H^{N}_{i}$ as derived from HNCO and HN(CA)CO peak-list

S.No.	$^{15}N_i$	$^{1}H^{N}_{i}$	$^{13}C'_{i}$	$^{13}C'_{i-1}$
1	111.65	7.81	174.92	177.86
2	111.84	7.84	174.92	175.08
3	111.87	8.76	174.80	175.54
4	112.05	7.85	175.24	178.01
5	112.12	7.53	174.92	176.57
.	.	.	.	.
.	.	.	.	.
169	134.55	8.09	175.57	177.10
170	135.37	7.94	178.50	175.94
171	135.54	8.29	175.71	176.72

2.8.8 Master_list

Master_list grouped and with 2-digit codes:

S.No.	$^{15}N_i$	$^{1}H^{N}_{i}$	$^{13}C^{\alpha}_{i}$	$^{13}C^{\beta}_{i}$	$^{13}C'_{i}$	$^{13}C^{\alpha}_{i-1}$	$^{13}C^{\beta}_{i-1}$	$^{13}C'_{i-1}$	code
1	114.64	7.98	51.81	20.25	176.62	50.55	37.53	178.45	2 0
.	.	.	.	.	.	.	.	.	
18	111.46	7.83	47.15	0.00	174.92	53.04	39.60	177.86	1 0
.	.	.	.	.	.	.	.	.	
32	115.69	7.63	61.42	63.74	175.12	59.21	33.15	178.21	3 0
.	.	.	.	.	.	.	.	.	
83	120.93	7.69	54.55	29.44	174.90	54.55	32.37	174.34	4 0
.	.	.	.	.	.	.	.	.	
122	122.73	7.80	65.94	28.76	177.51	59.49	28.76	178.77	4 1
.	.	.	.	.	.	.	.	.	
142	117.89	7.95	63.60	38.64	178.45	60.86	38.64	176.63	5 0
.	.	.	.	.	.	.	.	.	.
215	123.65	8.26	66.21	38.64	177.13	57.97	42.60	179.55	5 1

2.8.8. pml.input

```
INPUT FILE FOR PREPARING MASTER_LIST

Please read the manual to understand each term below

PEAK LIST FILENAMES:

cbcanh filename:            cbcanh.pl
cbcaconh filename:          cbcaconh.pl
hnco/hn(coca)ha filename:   hnco.pl
hncaco/hn(ca)ha filename:   hncaco.pl

TOLERANCE LIMITS:

Windows for peak selection along 1H, 13C and 15N dimensions
(peaks outside this limits are filtered out from the peak lists)

N15:             110 ppm to  136 ppm
1H(amide):       6.5 ppm to 11.0 ppm
1H(alpha):       3.5 ppm to  6.0 ppm
13C(alpha/beta):  15 ppm to 72.5 ppm
(13C-carbonyl are not required to be filtered)

Tolerance limits below this line need not be changed unless required

Tolerance limits for peak grouping:

(1) CBCANH spectrum

(a) For identification of CBCA(CO)NH peaks in CBCANH:

Min tolerance along N15 dimension: 0.50  ppm
```

```
Max tolerance along N15 dimension: 1.50  ppm

Min tolerance along 1H dimension: 0.05 ppm
Max tolerance along 1H dimension: 0.10 ppm

Min tolerance along 13C dimension: 1.00 ppm
Max tolerance along 13C dimension: 2.25 ppm

(b) For Searching self peaks in CBCANH spectrum:

Min tolerance along N15 dimension: 0.20 ppm
Max tolerance along N15 dimension: 0.575 ppm

Min tolerance along 1H dimension: 0.02 ppm
Max tolerance along 1H dimension: 0.07 ppm

(2) CBCACONH spectrum

Identification of sequential CA and CB peaks for a specific pair of N
and HN chemical shifts:

Min tolerance along N15 dimension: 0.10 ppm
Max tolerance along N15 dimension: 0.225 ppm

Min tolerance along NH dimension: 0.02 ppm
Max tolerance along NH dimension: 0.12 ppm

(3) HN(CA)CO and HNCO spectrum

Identification of CO peaks for a specific pair of N and HN chemical
shifts:

Tolerance along N15 dimension: 0.20 ppm
Tolerance along NH dimension:  0.02 ppm

Identification of self and sequential CO peaks for a specific pair of
N and HN chemical shifts:

Min tolerance along N15 dimension: 0.10 ppm
Max tolerance along N15 dimension: 0.65 ppm

Min tolerance along NH dimension: 0.020 ppm
Max tolerance along NH dimension: 0.095 ppm

(4) For preparation of master_list:

Min tolerance along N15 dimension: 0.20 ppm
Max tolerance along N15 dimension: 0.75 ppm

Min tolerance along NH dimension: 0.020 ppm
Max tolerance along NH dimension: 0.085 ppm
```

2.8.9 pps

```
1     M
2     A
3     E
```

```
4     A
5     L
6     F
7     K
8     E
9     I
.     .
.     .
.     .
130   L
131   E
132   F
133   S
134   L
```

2.8.10. assign.input

```
INPUT FILE FOR ASSIGNMENTS

Please read the manual to understand each term below.

Filename containing protein primary sequence: pps
Fileneme containing unassigned chemical shifts:unassigned_master_list
Filename containing unassigned residues:unassigned_pps

Minimum tolerance required for C-alpha search:   0.50 ppm
Maximum tolerance required for C-alpha search:   1.50 ppm

Minimum tolerance required for C-beta search:    0.20 ppm
Maximum tolerance required for C-beta search:    0.50 ppm

Minimum tolerance required for C-carbonyl search: 0.10 ppm
Maximum tolerance required for C-carbonyl search: 0.125 ppm

Maximim number of mismatches to be allowed in any stretch:  3
Minimum number of residues to be assigned in any stretch:   6

Number of residues after which tolerance will be increased: 50

Minimum tolerance required for C-alpha search:   0.50 ppm
Maximum tolerance required for C-alpha search:   2.50 ppm

Minimum tolerance required for C-beta search:    0.20 ppm
Maximum tolerance required for C-beta search:    0.80 ppm

Minimum tolerance required for C-carbonyl search:   0.15 ppm
Maximum tolerance required for C-carbonyl search:   0.25 ppm

Maximim number of mismatches to be allowed in any stretch:  2
Minimum number of residues to be assigned in any stretch:   2
```

(The tolerance limits in the above input file can be altered to check the unambiguity of the assignments)

2.8.11. assignment.list

```
  4 A 128.26       8.61 54.13 19.01 179.18      56.89 30.27 176.72
  5 L 122.96       8.29 57.44 42.33 177.61      54.13 19.01 179.18
  6 F 120.94       7.49 62.23 39.32 176.59      57.44 42.33 177.61
  7 K 117.39       7.48 58.80 32.59 177.95      62.23 39.32 176.59
  8 E 120.25       7.49 58.40 29.99 177.79      58.80 32.59 177.95
 10 D 126.46       8.57 61.28 37.81 176.23      51.67 19.03 177.96
 11 V 132.99       9.45 57.02 33.56 176.93      61.40 37.95 176.19
 12 N 122.80       8.79 63.20 40.42 176.67      57.16 33.70 177.10
 13 G 111.46       7.83 47.15  0.00 174.92      53.04 39.60 177.86
 14 D 121.54       8.12 53.33 40.43 177.83      47.55  0.00 174.91
 15 G 115.85      10.37 45.91  0.00 172.59      53.45 40.41 177.74
 16 A 125.21       8.00 50.43 22.59 175.80      45.92  0.00 172.54
   . .  .            .     .     .      .          .     .      .
   . .  .            .     .     .      .          .     .      .
   . .  .            .     .     .      .          .     .      .
128 E 120.92       7.59 58.95 30.27 179.75      60.73 29.16 180.03
129 F 124.62       8.74 61.41 40.69 176.53      58.95 30.27 179.75
130 L 116.62       8.31 55.76 40.97 178.04      61.41 40.69 176.53
131 E 119.36       7.39 55.93 30.12 176.58      55.76 40.97 178.04
132 F 125.36       7.81 57.98 40.42 174.04      55.93 30.12 176.58
133 S 121.73       6.98 57.30 64.56 172.05      57.98 40.42 174.04
134 L 133.13       7.44 57.03 43.72 181.76      57.30 64.56 172.05
```

2.8.12. Example run of program "assign"

```
$ assign
Are you assigning the first polypeptide stretch? (y/n):y
Number of amino acid residues to be assigned in this run:30

116 A to 129 F assigned
24 A to 33 K assigned
12 N to 17 V assigned

Total number of residues assigned so far = 30

Done

$ assign
Are you assigning the first polypeptide stretch? (y/n):n
Number of amino acid residues to be assigned in this run:30

45 I to 55 Q assigned
88 G to 96 E assigned
80 L to 87 D assigned
130 L to 134 L assigned

Total number of residues assigned so far = 63

Done

$
```

==========0==========

Practicals on

NMR Protein Structure Calculation Using CYANA

Peter Güntert
RIKEN Genomic Sciences Center, Yokohama, Japan

October 2003

Contents

Objectives

In these practicals one can learn to use the program CYANA for NMR protein structure calculations to

- run a conventional structure calculations based on manual NOESY assignment
- run combined automated NOESY assignment and structure calculation
- analyze and compare the results of structure calculations with conventional and automated NOESY assignment
- visualize and analyze protein structures with the program MOLMOL

Introduction: Data Files and Programs

The practicals are based on experimental NMR data for the hypothetical rhodanese domain At4g01050 from *Arabidopsis thaliana*. It is a 134-residue protein.

Data Files

In the top directory for these practicals one can find the following subdirectories:

Group*N*: working directory for group number *N* = 1, 2,…

Practical1	working directory for Practical 1
Practical2	working directory for Practical 2

bin: executable programs

cyana	CYANA version 2.0
cyanajob	script to start parallel CYANA structure calculations
cyanatable	script to produce overview table for CYANA calculation
cyanafilter	script to analyze results from automated NOESY assignment
cyanaclean	script to clean up a directory from old CYANA output files
Molmol	MOLMOL with automatic coloring of structure bundles
molmol	MOLMOL, standard version

Programs

These practicals use the new version 2.0 of the program CYANA that includes the DYANA torsion angle dynamics algorithm for the structure calculation[3] and a new algorithm for automated NOESY cross peak assignment similar to the CANDID method[2]. In addition, the program MOLMOL[5] is used for the visualization and analysis of the three-dimensional structures.

Start

To start with the practicals, change to the subdirectory Group*N*.

Please do not modify any files outside your Group*N* directory and its subdirectories.

Practical 1 – Conventional CYANA Structure Calculation

The aim of this practical is to learn how to perform a conventional structure calculation using the program CYANA with assigned NOESY peak lists and to analyze its output files. The following input data files are provided in the subdirectory 'Group*N*/Practical1':

atr.seq amino acid sequence

atr.prot chemical shift list

init.cya initialization macro (script for CYANA) that will be executed automatically when starting the program CYANA

CALC.cya CYANA macro for structure calculation based on assigned NOESY peak lists

NOESY15N.peaks
assigned peak list in XEASY format from a 3D ^{15}N-edited HSQC-NOESY

NOESY13C.peaks
assigned peak list from the aliphatic region of a 3D ^{13}N-edited HSQC-NOESY

NOESY13CARO.peaks
assigned peak list from the aromatic region of a 3D ^{13}N-edited HSQC-NOESY

To start the structure calculation on your workstation, you first start the program CYANA by typing 'cyana'. It will automatically read in the initialization file 'init.cya'. Type 'CALC' at the CYANA prompt to execute the macro 'CALC.cya'. First a report of the possible inconsistencies between the peak lists and the chemical shift list will be displayed on the screen. Then the actual structure calculation will start. This calculation can be done in a few minutes. After the calculation the following files will be present:

atr.upl list of upper limit constraints used during the structure calculation

atr.aco list of angle constraints used during the structure calculation

dcostat.ps plot of the number of distance constraints per residue in Postscript format

ramaplot.psRamachandran plot in Postscript format

atr.ovw overview of the CYANA target function, violated constraints and RMSD of the 20 structures with the lowest CYANA target function.

atr.cor coordinate file containing the 20 conformers with the lowest CYANA target function.

The Linux command 'gs' can be used to display the postscript files 'dcostat.ps' and 'ramaplot.ps'. The structures can be visualized using MOLMOL with the command 'molmol -r 7-125 atr.cor' given at the Unix prompt. The '-r 7-125' option is used to superimpose the individual conformers in the structure 'atr.cor' for minimal RMSD of the backbone atoms of residues 7-125.

Practical 2 – Automated CYANA Structure Calculation

The aim of this practical is to perform a complete structure calculation with fully automatic NOESY assignment, starting from unassigned NOESY peak lists in the from the spectrum analysis program NMRView. The program CYANA is used for automated combined NOESY assignment and structure calculation. Finally, the results of the automated calculation are analyzed in the context of summary statistics, and the three-

dimensional structures. The structures obtained by the conventional and automated approach will be compared, too. The files for this practical are located in the subdirectory 'Group*N*/Practical2':

noesy15N.xpk
: 3D ^{15}N-edited HSQC-NOESY peak list in the format of NMRView

noesy13C.xpk
: 3D ^{13}C-edited HSQC-NOESY peak list (aliphatic region)

13Chsqcaro.xpk
: 2D ^{13}C-HSQC peak list (aromatic region)

15Nhsqc.xpk
: 2D ^{15}N-HSQC peak list

atr.seq amino acid sequence

atr.prot chemical shift list

init.cya initialization macro for CYANA

CALC.cya CYANA macro for automated structure calculation based on unassigned NOESY peak lists

Step 1: Automatic NOESY assignment and structure calculation

To perform the automated NOESY cross peak assignment and structure calculation, edit the 'CALC.cya' macro by replacing the file name 'testint' with the name of the newly created peak list. The complete CYANA run consists of 7 cycles of automated NOESY assignment and structure calculation and a final structure calculation and would be slightly time-consuming when using a single processor. However, it can be run efficiently on a parallel computer (ask the instructors for details) using the 'cyanajob' command:

```
cyanajob –n M CALC
```

where *M* is the number of processors to use in parallel and CALC the name of the macro file to execute, 'CALC.cya'. During the calculation, the output from all processes will be collected in the 'CALC.out' file. To read this file in an interactive way you can use the Unix command 'tail –f CALC.out'. In each cycle 1,…,7 a NOESY assignment file (*.noa), an upper distance limit file (*.upl), a structure coordinate file (*.cor), and an overview file (*.ovw) will be written. The results of the final structure calculation will be stored in the files 'final.*'.

Step 2: Analysis of the results in terms of statistics and 3D structure

The command 'cyanatable' generates an overview table with statistical information on the number of peaks, upper distance limits, target function values and RMSDs that have been obtained in the cycles 1-7 of automated NOESY assignment and structure calculation, and in the final structure calculation.

The command 'cyanafilter' can be used to extract information on particular peaks, or classes of peaks from the generally large NOESY assignment (*.noa) files according to various criteria, e.g.

```
cyanafilter –u cycle2.noa          # information on peaks unassigned in cycle 2
cyanafilter –V 1.0 cycle7.noa      # peaks causing a constraint violation > 1 Å in cycle 7
```

More information about the 'cyanafilter' command can be obtained by the option '–h'.

Step 4: Comparison of structures from conventional and automated approaches

Now we will be able to compare the structures obtain in the convectional and automated structure calculation and also the structures from the intermediate cycles, using MOLMOL.

The command 'Molmol –r 7-125 final.cor ../Practical1/atr.cor' will superimpose the final structure from the automated approach on the structure obtained from manually assigned NOESY peaks lists in Practical 1.

The command 'Molmol –r 7-125 *.cor' will superimpose the structures from all cycles of the automated approach.

References

1. Güntert, P. (2003) Automated NMR structure calculation. *Prog. NMR Spectrosc.*, in press (available online).
2. Herrmann, T., Güntert, P. & Wüthrich, K. (2002). Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J. Mol. Biol.* **319**, 209-227.
3. Güntert, P., Mumenthaler, C. & Wüthrich, K. (1997).Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* **273**, 283-298.
4. Güntert, P., Mumenthaler, C. & Herrmann, T. (1998). DYANA 1.5 User's manual. http://cyana.gsc.riken.go.jp/Software/DyanaManual.pdf
5. Koradi, R., Billeter, M. & Wüthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51-55.

REDCAT: A New Residual Dipolar Coupling Analysis Software Tool

Input Format

Input to REDCAT consists of nine columns of data space or tab delimited for every entry. The nine columns are: x, y and z coordinates of the first atom, x, y and z coordinates of the second atom, experimental RDC, maximum RDC for that vector type at 1 A distance and measurement error. You can refer to "1a1z.redcat" that is included with this package as an example. The following is an example input line:

-7.679 10.035 5.255 -6.713 10.289 5.296 14.789 24350 1 /*from 2*/

Note that any information after the 9th column is considered to be comment. For backward compatibility, you may choose to start and end the comment section with /* and */ (just like commenting in C).

The input file can either be prepared manually or using some of the other scripts that are under development. Perl scripts "MakeREDCAT.prl" and "MakeREDCAT_chain.prl" are currently the only two programs that are included. The only difference between these two is that the second one will extract the information only from the appropriate unit in a multi-subunit molecule. Both of these programs prepare a REDCAT input file from a given PDB file. The user is then required to manually edit the file to include the RDC and error values. The usage of these programs is as following:

MakeREDCAT.prl "first atom" "second atom" "MaxRDC"

MakeREDCAT_chain.prl "first atom" "second atom" "MaxRDC" "chain"

Standard input and output are the assumed input/output sources. Use redirection to feed an input or capturing the output into a file.

Example: To prepare a file which contains the coordinates for amide (atom name N) and amide-proton (atom name H in our PDB file) the following command will be issued:

MakeREDCAT.prl N H 24350 < input_file.pdb > output_file.redcat

In preparing your input file please ensure that the following criteria are met:

1. There should be on leading spaces in the input files.
2. Each datum in the input file needs to be separated by only one delimiter. Do not put multiple spaces or tabs in the input file to make it esthetically pleasing to the eye.
3. There should be no blank lines at the beginning, in the middle or at the end of the input file.

The following values of RDC have special meaning and are reserved for specific use:

1. A RDC value of 999 is interpreted as a missing value and will automatically be excluded from all order tensor solution analyses (refer to the next section).
2. A RDC value of "AVG" will indicate and averaging of the RDC values (This feature will be available in the next release scheduled on 8/20/03).

Main Graphical User Interface (GUI)

Figure 1 below is a snapshot of the initial screen after having loaded an input file. This GUI displays all of the relevant information from the input file. The only field that is hidden from the user is the maximum RDC. The remaining fields that have been displayed are subject to change from within this interface. The select buttons shown on the left hand side of this figure allow for the inclusion/exclusion of a particular entry in the analysis. A value of 999 for RDC will automatically causes and exclusion of that entry in the analysis.

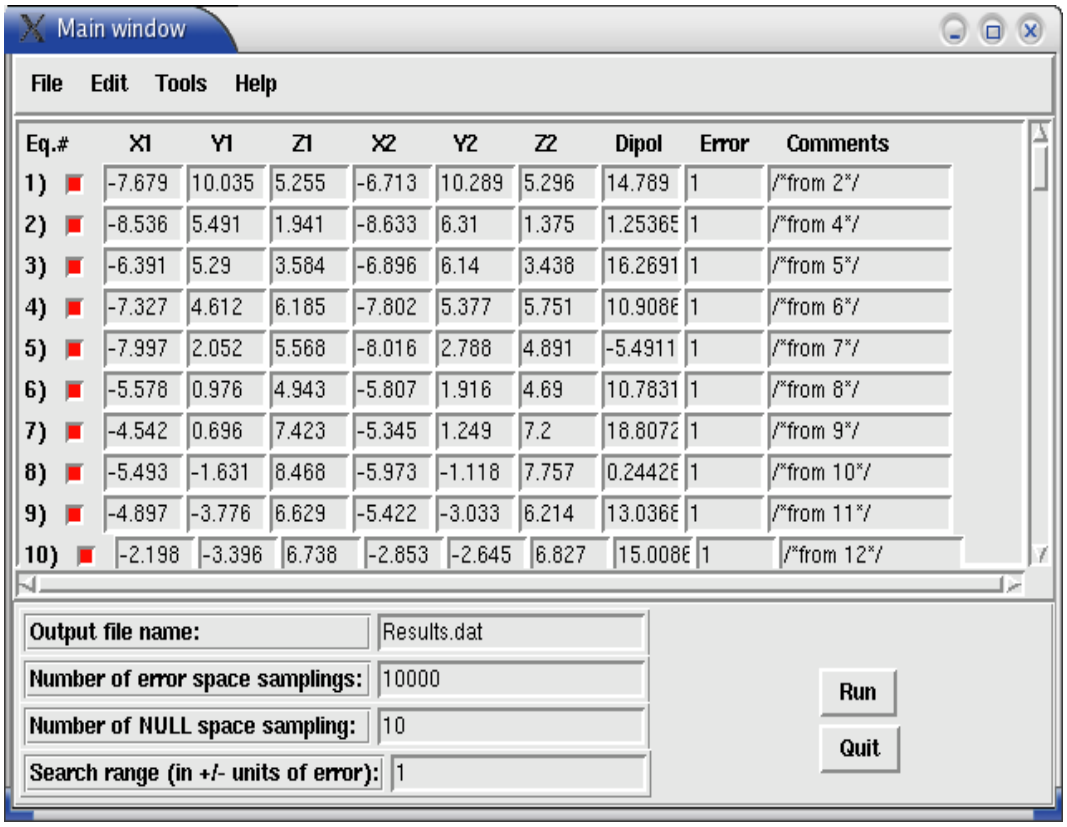


Figure 1. Main Graphical User Interface.

The following is a brief description of the input fields in the main GUI:

1. *Output file name:* Allows the user to capture the result of the analysis into an output file. This value by default is chosen to be “Results.dat” and can be changed by the user.

2. *Number of error space samplings:* This value determines the number of times the error space is to be sampled. In short, this is the number of times the experimentally measured RDCs will be randomly regenerated in the error range indicated. For more detailed description of this entry please refer to our publication soon to appear in JMR.

3. *Number of NULL space sampling:* Under the conditions where the system of study is ill-conditioned, null-space contributions to the order tensors can be accommodated. This value indicates the number of times random contributions from the null space should be added to each of the solutions. In the above case the null space will be sampled 10 times for each of the 10000 instances of the RDC for a total sampling of 100,000 times. For more detailed description of this entry please refer to our publication soon to appear in JMR.

4. *Search range (in +/- units of error):* Under some special conditions, it is possible to sample a point outside of the range of error such that it reproduces acceptable results (reproduces back calculated RDCs within error). While this phenomenon is both theoretically possible and unnecessary simultaneously, it may provide some useful functions. We suggest that only advance users of this program alter this value to other than 1. This parameter may just simply overcome some sampling properties of the solution space.

5. *RUN:* This button will engage the analysis of the data. Upon clicking of this button certain number of programs will engage and produce the final results in a pipeline fashion. These commands can be evoked from the command prompt manually or by an automated script. For more detailed description of the analysis procedure please refer to our publication soon to appear in JMR. There will be more included in this manual at a later time.

6. *Results/Message Window:* Results of any analysis is displayed within the "Message!" window as demonstrated in Figure 2. After clicking on the "RUN" button, general information regarding the status of analysis are displayed in this window. The following information will be listed respectively:

    - *Rank of the system*. If the system is ill-conditioned (rank less than 5), the effective rank of the system will be displayed here. The effective rank of the matrix is determined as a function of the numerical precision of the computer. This feature is likely to change in the future to be a measure of the indicated errors and not the numerical precision of the machine.

    - *Rejection Status:* Contribution of each entry to the rejection of order tensor solutions is listed here. Any one entry that causes maximum number of rejections (for example 10000 out of 10000) is indicated in red, otherwise it is indicated in green or gray. Green entries are the ones that did not cause 100% rejection and gray entries are the ones that are excluded from the analysis. Keep in mind that all entries can appear green but over all have a 100% rejection of the trials. This event takes place due to the overlapping of the rejections. Analysis of these data under simple circumstances can be informative. For example if only one of the entries

produces 100% of rejections, then it is very likely that the RDC and the coordinates of that vector are sever violation with respect to the remainder of the entries. Exclusion of this entry should produce solutions.

Other Information such as the rejection status, order parameter solutions, the corresponding Euler angles, best solution and error analysis will all be concatenate to the contents of this window. This information can be saved into a file by using the save menu option in this window.

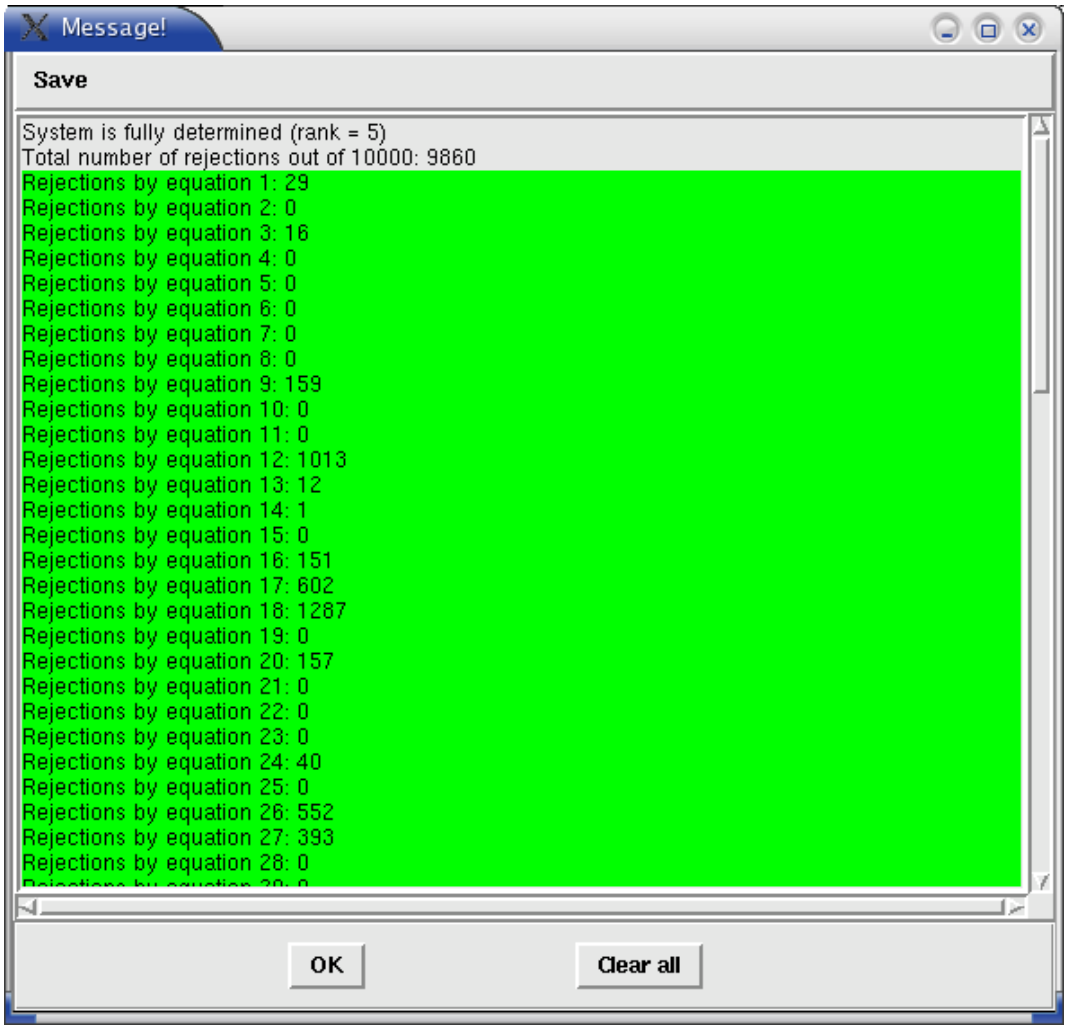


Figure 2. An example of Message! Window displaying analysis status.

7. *Quit:* This button will quit out of the program.

File Menu

What can I say? Nothing to say here. Most things are like anything else. More options are apt to appear here at some time in the future.

Edit Menu

Most elements of this menu are self explanatory. The “Options” menu is a blank page at this point; however the near future, this will be the location where intermediate file names and other features will be tailored.

Tools Menu

- **Get Solutions:** After a successful analysis of the data (by pressing the Run button), this option will retrieve and display the solutions as shown in Figure 3 below. Here S_{xx}, S_{yy} and S_{zz} are the three principle order parameters, a, b and c are three Euler angles relating the molecular frame to the principle alignment frame (PAF), Eta is related to the rhombicity of the alignment (by a factor of 2/3) and GDO is the general degree of order. For more detailed description of these entities please refer to our publication soon to appear in JMR.

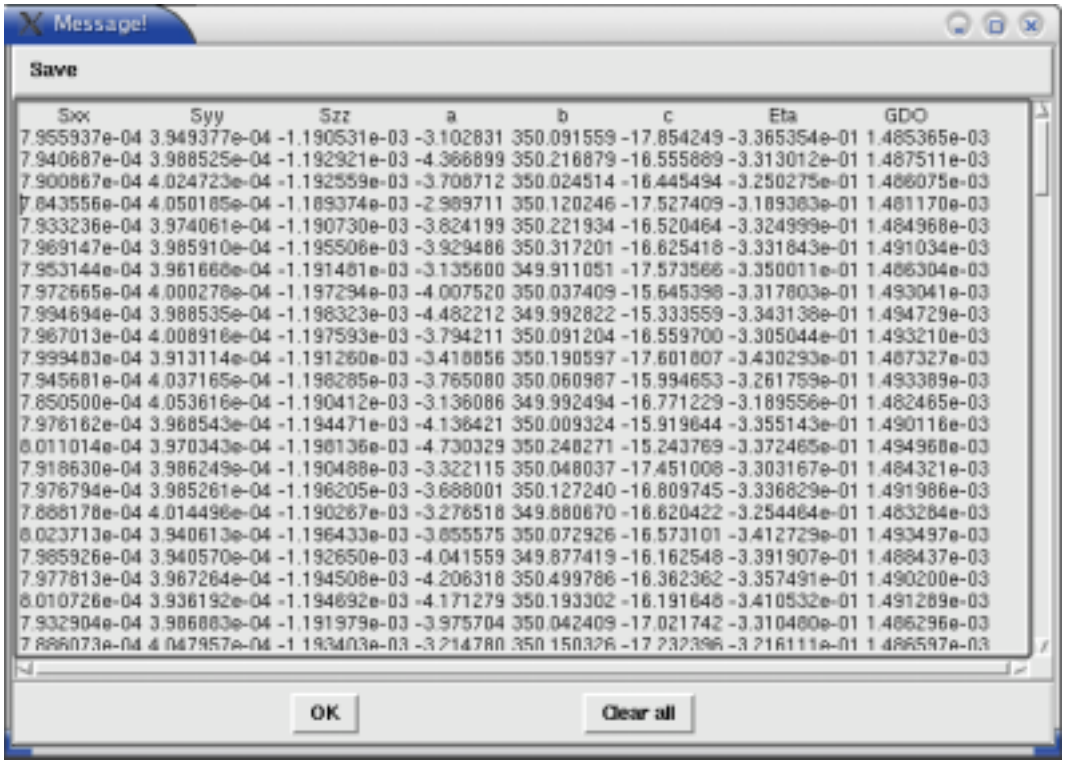


Figure 3. List of solutions reported by Tools:Get Solutions.

- **Get Best Solution:** This option will retrieve the best solution order tensor as shown in Figure 3. This is the order tensor that will back calculate the RDC data with the minimum rmsd. It is possible for this option to return no solution even though there should always exist a best solution to any problem (in this case, regardless of the degree of agreement between the data and structure). If the best solution does not reproduce back calculated RDCs within error of the experimental data, this procedure will not report any solutions. This is only to prevent the novice users from miss interpreting the results. In general one can expand the errors to very large numbers and always get the best solution. Also, one can perform the Error Analysis in order to get the errors required for the attainment of the best solution. Another thing that users of the above two functions should be aware of is that any of the four following events is possible.

    1. Both "Get Solutions" and "Get Best Solution" will return results.
    2. "Get Solutions" returns results but "Get Best Solution" does not.
    3. "Get Solutions" does not display any results, but the "Get Best Solution" displays resutlts.
    4. Neither one gives any results.

Sampling properties and other characteristics of singular value decomposition can be used to explain any of the above four cases.

- **Calculate/Substitute RDC:** This tool will display the interface shown in Figure 4 below and allows the back calculation of RDC with a given set of S_{xx}, S_{yy}, S_{zz}, a, b and c. The "Error" field will allow for the added noise of indicated range. For example, error of 2 Hz would add uniformly distributed noise within the range ±2 Hz. Once the "Calculate RDC" button is clicked the back calculated RDCs will be displayed in the text box below. At the bottom of this page the rmsd between the back calculated and measured data will be displayed. If the " Substitute RDC" check button has been selected, then the newly calculated RDCs will substitute the measured data in the main window. The substitute option will substitute all entries regardless of their selection status.

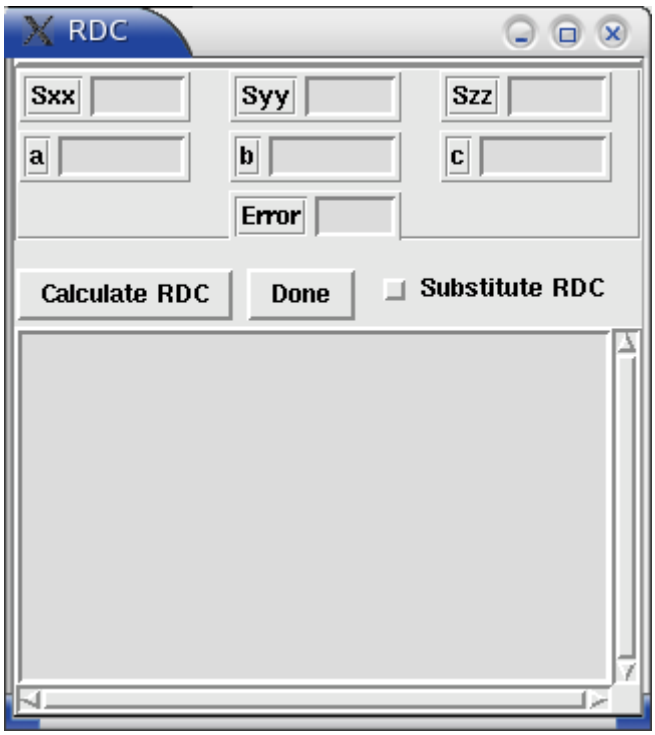


Figure 4. Calculate/Substitute RDC interface.

- **Perform Error Analysis:** It is likely that the analysis of the empirical data does not produce any solutions. This is due to internal inconsistencies between the collected RDC data and the structural information. Furthermore, moderate to severe inconsistency of one vector could cause 100% rejection by several other vectors. This complicates the identification of the faulty data. Conventionally, the user will have to engage in random alteration of errors in combinatorial fashion. This effort is often fruitless and tedious at best. Error Analysis is a tool that aims to make this task simpler. Figures 5 and 6 below show two examples of faulty data of different severities. Figure 5 is the rejection status when the first RDC has been altered by 2 Hz. In this case the first entry has clearly been identified as the problematic one and can therefore be corrected or excluded. Figure 6 lists the rejection status of the analysis when the first RDC has been negated in sign. In this case 15 other entires have also caused a 100% rejection rate besides the first one. Here the isolation of the problematic vector is practically impossible. The results of the "Error Analysis" shown in Figure 7 clearly identifies the faulty data.

In general, the error analysis will report the error values required to at least one solution. In our example from above, the examination of the errors reveals that the first entry will require the expansion of error to a value of 26.73. This value

becomes significant when compared to the remaining error values of less than approximately 2 Hz. The result of "Error Analysis" will indicate all of the entries which require and expansion of error in red while the remaining entries will be displayed in green or gray as before. Incorrect interpretation of these results can be very harmful since any scenario can be manipulated to produce an answer. The suggested expansion of errors need to be confirmed and justified by the user based on experimental conditions.

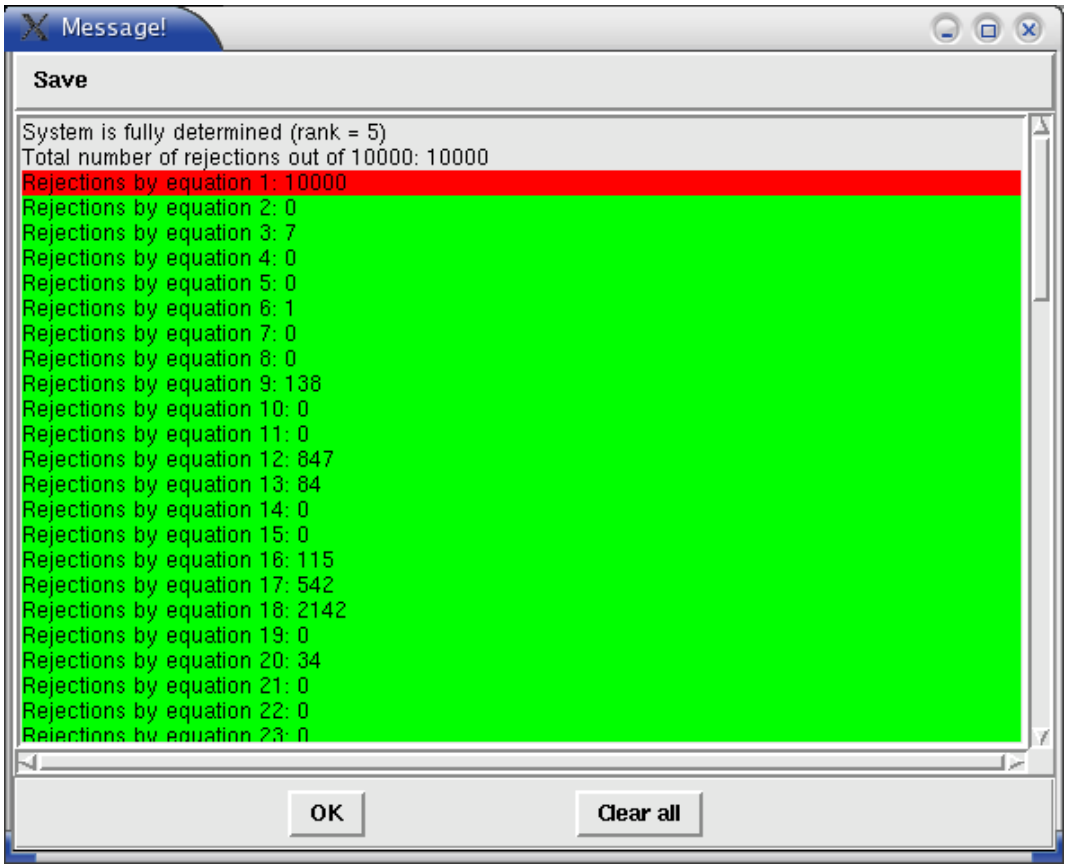


Figure 5. Rejection status for introduction of modest error (2Hz) into the first RDC.

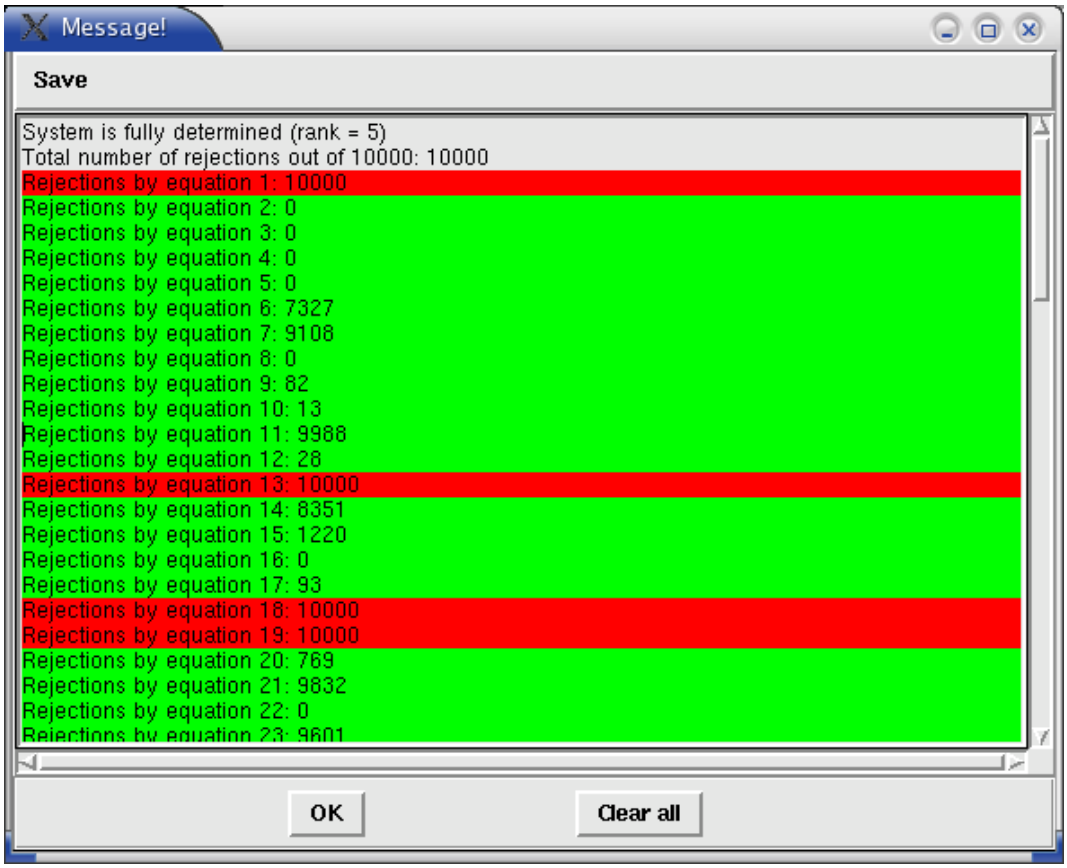


Figure 6. Sign reversal of the first RDC entry causes 100% rejection by 15 out of 80 other vectors.

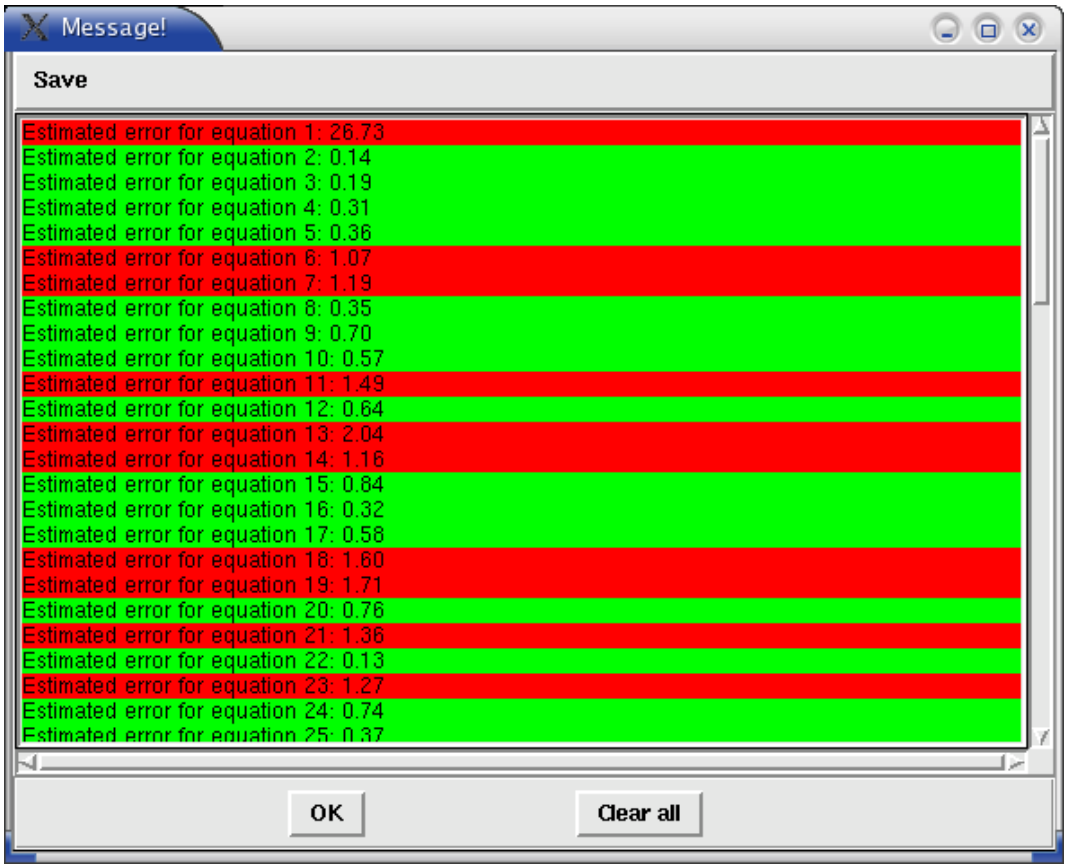


Figure 7. Sign reversal of the first RDC (from 14 Hz to -14 Hz) entry causes 100% rejection by 15 other vectors.

- **Get Estimated Errors:** As in the previous example, often times large number of error limits need to be modified in order to obtain a solution. Although an estimate for these errors can be obtained by the use of the Error Analysis function, manual alteration of all of these values can be tedious. Under such conditions the "Get Estimated Errors" can be utilized to update all errors based on the error analysis results. These errors are not exactly as indicated by the Error-Analysis. In fact they are increased by 0.1 Hz. This value is fixed in the current version, but will be included under the options menu in the future versions. The increased value in the error is to increase the size of the solution space for a more successful sampling procedure.

- **Rotate PDB:** Following a successful analysis session, it is beneficial to express all coordinates in the principle alignment frame. Rotate PDB will rotate a PDB using specified rotation angles. Figure 8 shows the basic interface for this function. This tool will rotate the coordinates only by performing Euler rotations. The angles of interest can come from the "Best Solution" or any of the list of solutions. The angles alpha, Beta and Gamma of

the interface correspond to a, b and c of the reported solutions respectively. When using this function, be mindful of the inversion properties of order tensors since any set of a, b and c can have 3 other related sets that are indistinguishable from one another on the basis of RDC data.

- **Rotate Coordinates:** The easiest way to align multiple fragments with respect to each other is to transfer all fragments into the principle alignment frame (PAF). Rotate PDB function of RDC can be used to perform just this task. Analysis of multiple fragments

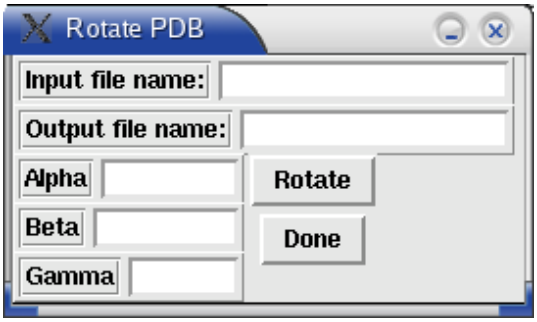


Figure 8. PDB rotation tool.

within the PAF will require the appropriate rotation of each PDB (for each fragment) followed by the preparation of REDCAT input file from each of these modified coordinates. The "Rotate Coordinates" tool of REDCAT produces this final result by simply rotating all of the coordinates already loaded into REDCAT. Figure 9 below shows the user interface for this function. The method of rotation can be chosen to be Euler rotation or rotation about an axis. Since the modification of coordinates applies only to those entries that are selected, the modification of the coordinates can be restricted to only a sub selection of the entries. This segment can then be analyzed separately in order to optimize the overall agreement of data. The Euler rotation requires three angles of rotation (a, b and c) while rotation about an axis requires the coordinates for the rotor and the angle of rotation about this axis.

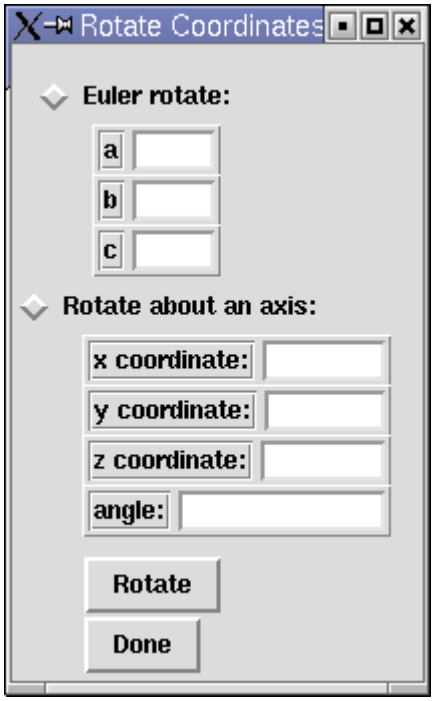


Figure 9. GUI for "Rotate Coordinates" function of REDCAT.

Note that here the rotor vector does not need to be normalized, this adjustment will be automatically performed. The x, y and z coordinates can be obtained by finding the x, y and z difference between two points that define the vector. Users are advised to be aware that rotation of θ degrees about a vector defined from point a to b is equivalent to -θ when the vector is defined between points b and a.

- **Dynamic Averaging:** Study of RDC can provide a great deal of information regarding the motions present in a molecule. Dynamic Averaging is a tool that allows the simulation of the effect of motion on the observed RDCs. The user interface for this tool is shown in Figure 10. Each of the required fields is described below.

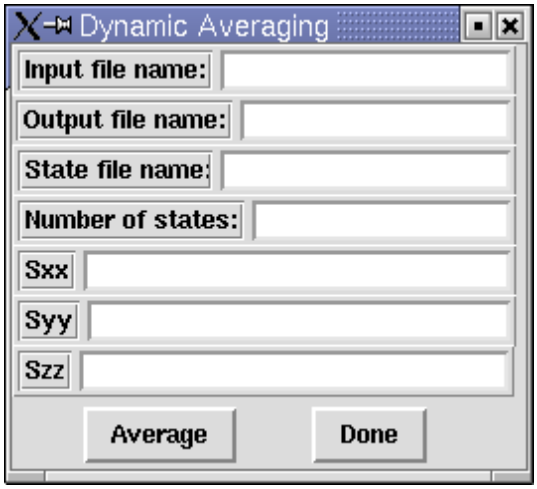


Figure 10. User interface to the Dynamic Averaging tool.

Input file name: Is the name of the input file to the Dynamic averaging calculations. This file needs to be in the REDCAT input format.

Output file name: This file will contain the results of the Dynamic Averaging calculations in the REDAT format. Since this file is in the proper format, it is already to be an input to the REDCAT program. Also keep in mind that the RDC of selected vectors will be replaced with the new RDC.

State file name: All of the necessary information such as description of different states and the fraction of populations in each state needs to be described in an input file. The name of the file that contains all that information needs to be inputted here. The format of this file will be discussed later.

Number of states: This number indicates the number of discrete states that are used to describe the motion. There should be this number of entries in the states file.

S_{xx}, S_{yy} *and* S_{zz}*:* Are the three principle order parameters (therefore all coordinates need to be in the PAF) that are reported by the static region of the molecule. This region of the molecule is the point of reference by which the internal motion is defined.

The states file is the location where each discrete state of motion has been defined. Each line of this file will consist of the description of one of the states. Therefore there need to be exactly the same number of lines in this file as indicated in the

"Number of states" field above. Each line will start with a leading 0 or 1. 0 is indicative of rotation about an axis while 1 indicates an Euler rotation. The last entry in each line indicates the fraction of population in that state. Note that the program does not check to see if the sum of all fractions equal to 1. this condition needs to be taken care of by the user. The intermediate entries in each line will depend on the type of rotation. If rotation about an axis is defined (leading 0), then four parameters namely x, y and z coordinates of the rotor axis and theta, the rotation about the rotor (in degrees) need to follow. If an Euler rotation is the defined method of rotation (leading 1), then a, b and c (the three Euler angles in degrees) need to be indicated. The states file can be any combination of Euler or rotor rotations in any order. Refer to the file "states.in" that is included with this package as an example. This file describes a 3 state jump that are related by 120 degrees about the rotor (1, 2, 3).

- **Sauson-Flamsteed Projection:** This tool creates a two dimensional projection of a globe that allows the visualization of the PAF with respect to the molecular frame. This tool takes advantage of the plotting program "gnuplot" and conversion program "convert" that are normally installed by default in most flavors of Linux. Aside from the availability of the above two programs, the file "map.out" needs to also be present in the directory where the analysis are being performed. This file is included in the REDCAT package.

Tutorial One:

Analysis of Simulated Data

Input File Generation

1. Generate the input file from 1A1Z protonated PDB file by typing the following command:

   MakeREDCAT.prl N H 24350 < 1a1z.H.pdb > ~/test.redcat

   Compare this file with 1a1z.redcat to make sure that you everything is working right. These two files should be identical with the exception of the RDC and error columns.

2. Start the program by typing REDCAT.tcl or whatever the link name is.
3. Load the input file by:
    1. Select “Load” from the File menu.
    2. Select test.redcat or type the name in the text field. Note that all RDC values are -1.0.
4. To simulate data:
    1. Select “Calculate/Substitute RDC” from the Tools menu.
    2. Here type Sxx = 0.0007, Syy = 0.0003, Szz = -0.001 (traceless), a = 5, b = 10, c = 15 and Error = 0.
    3. Check the “Substitute RDC” box and click on the “Calculate RDC” button. This should replace all of the RDCs in the main window. You should see a windows similar to Figure 11 below.

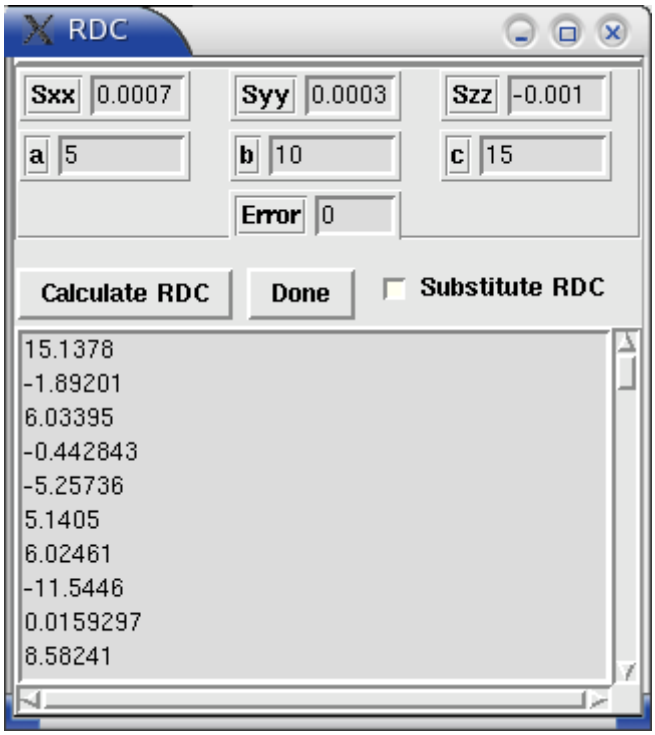


Figure 11.

5. Engage the analysis engine by clicking on “Run” in the main window.
6. The “Message!” window should appear after a few seconds (depending on the computational resources) with the analysis status that should appear like

Figure 12. You should have near 0 rejections due to the perfect nature of the data.

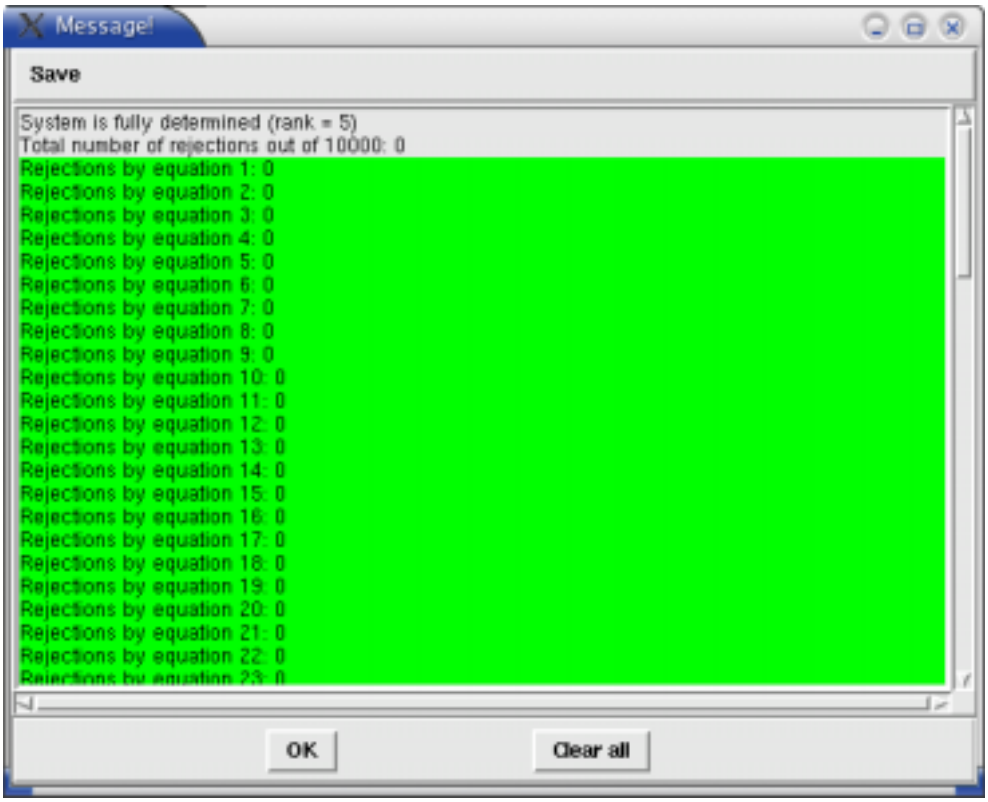


Figure 12.

7. To observe the solutions, use the “Get Solutions” or “Get Best Solution” options in the Tools menu. You can clear the contents of the “Message!” window by using the “Clear all” button or saving them by using the save option in this window. You can now use any of the solutions in order to back calculate RDCs to confirm correct function. Also you can use the rmsd value reported by the “Calculate/Subs...” tool to confirm that the best solution does result the best rmsd value.

8. For visual inspection of the PAF within the molecular frame select the “Sauson-Flamsteed Projection” from the Tools menu. You should see the same as in Figure 13. If the program returns an error message related to “map.out” make sure that you copy the file map.out from the install directory to your working directory. Repeat this procedure twice again. The first time may not display the plot! This is a bug that will be fixed soon.

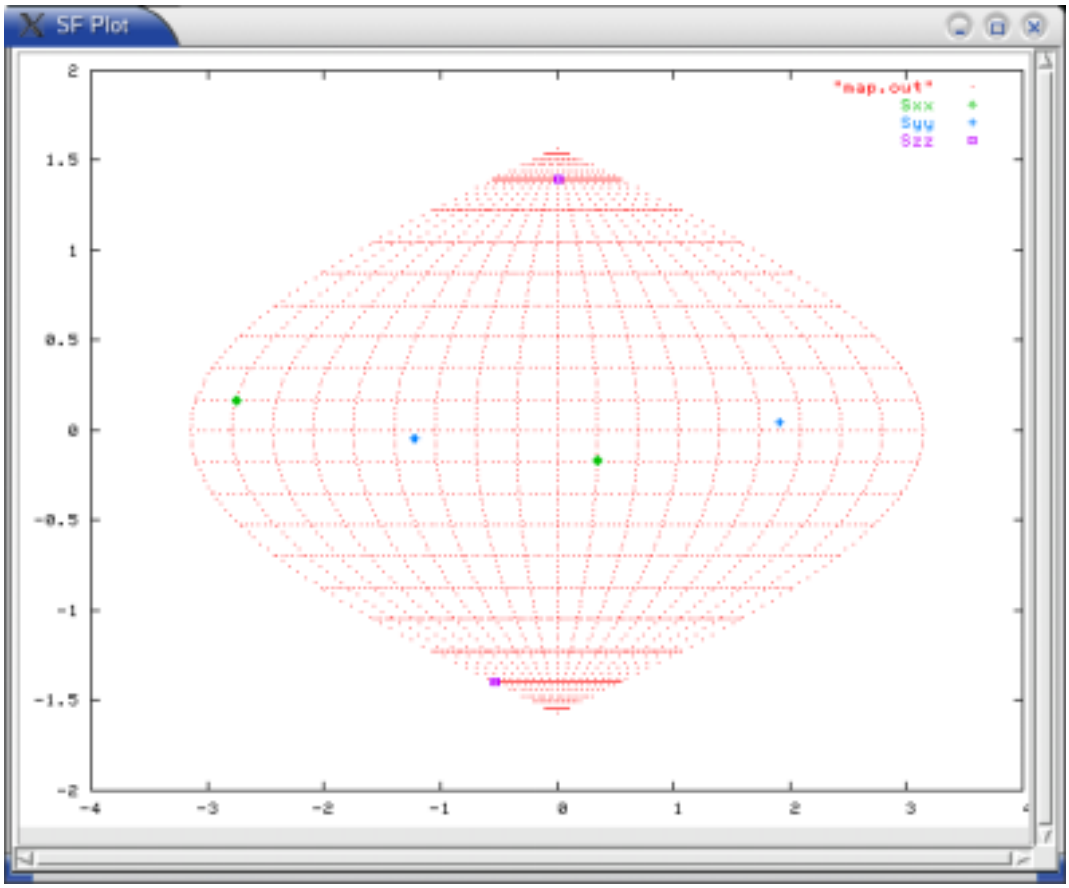


Figure 13

Tutorial Two:

Analysis of Simulated Data with Error

1. Using the input file prepared from the first tutorial, generate RDC values with the values shown in Figure 14. Note that due to the state of your random number gnerator, it is likely that you will not get the same data as shown below.

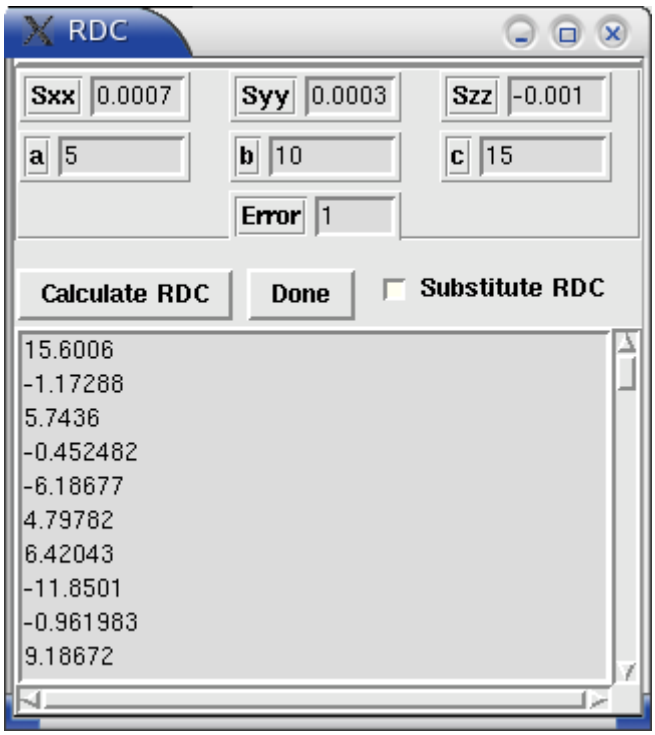


Figure 14

2. Run the analysis and note the results. Most likely you should get a large list of red entries (indicating 100% rejection). However, it is possible that you get solutions since it is possible that the random number generator did not produce any numbers larger than 0.2.

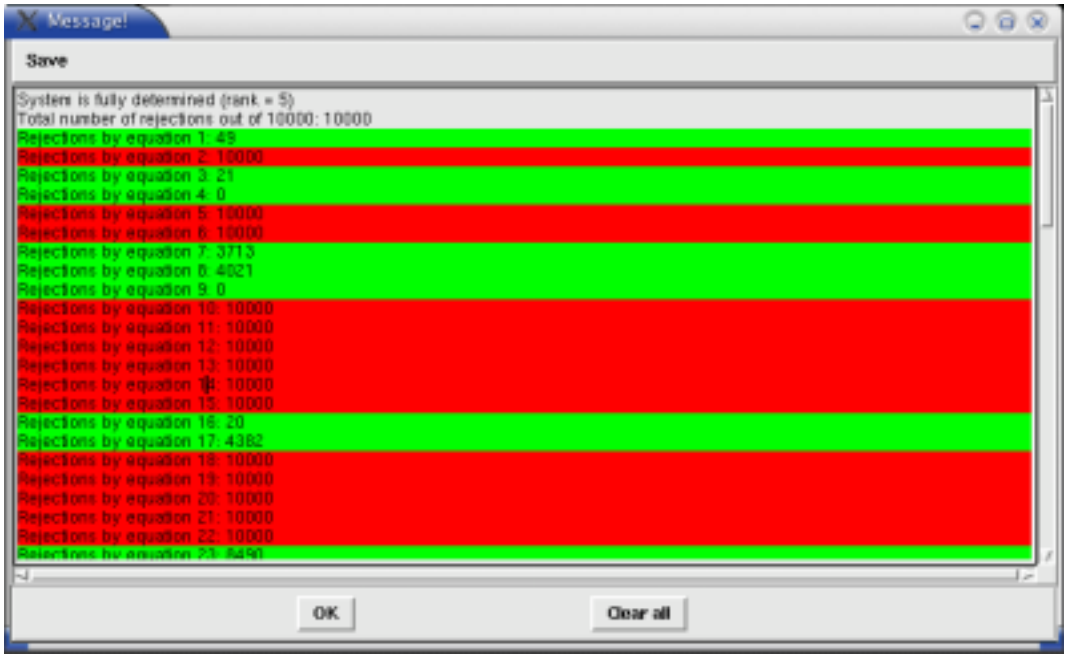


Figure 15

3. To identify the faulty entries select the “Perform Error Analysis” from the Tools menu.

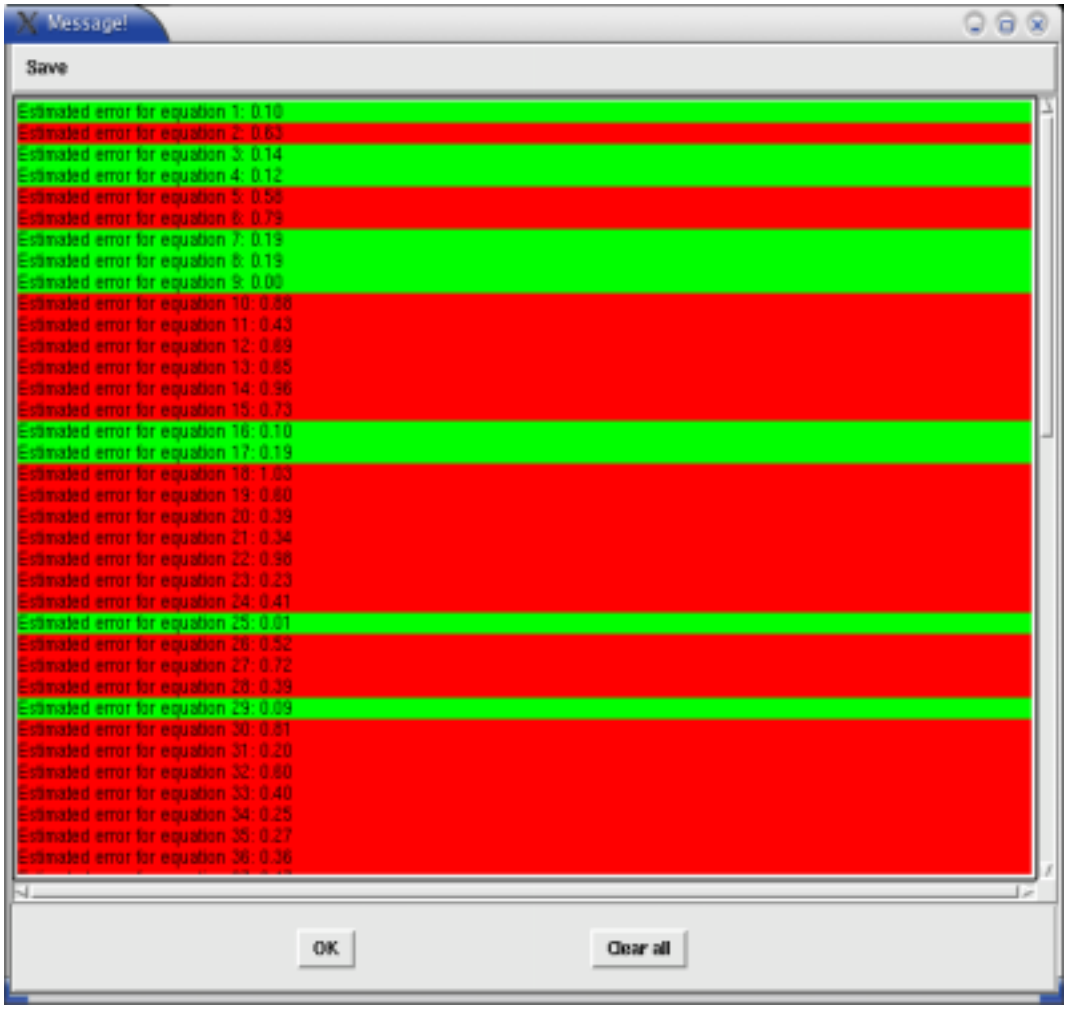


Figure 16

4. You can either manually correct each of the indicated errors or select “Get Estimated Errors” from the Tools menu. Selecting “Get Estimated Errors” will substitute all errors (including the green ones) with the suggested errors plus 0.1.

5. Perform the analysis again. Now you should at least have a best solution. You may or may not have a list of solutions.