Journal of Biomolecular NMR, 1 (1991) 217-236
ESCOM

J-Bio NMR 024

A relational database for sequence-specific protein NMR data

Beverly R. Seavey, Elizabeth A. Farr, William M. Westler and John L. Markley*

Biochemistry Department, College of Agricultural and Life Sciences, University of Wisconsin-Madison, 420 Henry Mall, Madison, WI 53706-1569, U.S.A.

Received 21 January 1991
Accepted 13 May 1991

Keywords: Peptides; Proteins; Experimental conditions; NMR spectral parameters; Structural, dynamic, kinetic, and thermodynamic results; Information retrieval

SUMMARY

A protein NMR database has been designed and is being implemented. The database is intended to contain solution NMR results from proteins and peptides (larger than 12 residues). A relational database format has been chosen that indexes data by: primary journal citation, molecular species, sequence-related and atom-specific assignments, and experimental conditions. At present, all data are entered from the primary refereed literature. Examples are given of sample queries to the database. Possible distribution formats are discussed.

INTRODUCTION

The methods for assigning NMR signals from biological macromolecules, especially proteins, have developed to the point where they have become relatively rapid and reliable (for reviews, see: Wüthrich, 1986; Markley, 1989; Stockman and Markley, 1990). The technology continues to evolve rapidly; recent developments are the steady publication of novel 3D and 4D NMR methods for spectral analysis (for example: Fesik et al., 1989; Zuiderweg and Fesik, 1989; Ikura et al., 1990; Kay et al., 1990). Advances in the methodology for assigning spectra, determining coupling constants, and measuring cross-relaxation have led to a proliferation of assigned primary NMR data (e.g., chemical shifts, coupling constants, relaxation and cross-relaxation rates) as well as derived data (e.g., 3D structures, chemical exchange rates, and pK_a values). The rate of publication of such data promises to accelerate as the new methodology is applied to the ever growing number of small natural proteins, mutant proteins, and protein fragments that are of in-

* To whom correspondence should be addressed.

terest to protein chemists and biologists. It is critical that NMR data from protein studies (such as assigned chemical shifts, input data for structural determinations, and derived structural coordinates) be readily accessible to the scientific community.

It is important that primary NMR data be published in conjunction with journal articles: such data are needed to evaluate the scientific conclusions, to support future comparative studies, and to avoid the necessity of repeating experiments. Journal editors are under increasing pressure to shorten papers by deleting such data. One solution is to provide the detailed data as supplementary material. Journal editors (and authors) have been criticized for publishing protein NMR papers in which key data, such as chemical shifts or NMR-derived structural constraints, were omitted or supplied only as supplementary material (Jardetzky, 1989). The Division of Biology and Medicine of the International Society of Magnetic Resonance (ISMAR) set up a task force to consider recommendations for the publication of biomolecular NMR data. Its recommendations (Division of Biology and Medicine, ISMAR, 1990) provide guidelines for data to be included in published articles and encourage the development of NMR databases.

A comprehensive repository for NMR-derived results on proteins will enhance the usefulness of such data to the community (Markley and Ulrich, 1984). Ideally, the data should be available in machine-readable form. This is preferable to looking up the data as scattered in various journals and their supplementary information. Substantial delays are involved in obtaining supplementary data by mail, and once obtained, this bulky material, which may consist of as many as 51 pages (Gao et al., 1990), may not be organized in a way that a given researcher finds useful.

We recently presented the rationale for a protein NMR database and provided a description of how it might be designed as a sequence-related database in flat-file format (Ulrich et al., 1989). With financial support from the National Library of Medicine, we have implemented a protein NMR database project (*BioMagResBank*). As the design of the database progressed, the advantages of using a relational database management system became apparent (Markley et al., 1990). We present here a description of the database design, a progress report on its development, and prospects for its distribution. The staff of the Protein Data Bank (*PDB*) independently have begun adding structural NMR data to its holdings. The aims of *PDB* and *BioMagResBank* are different and complementary; and, as described below, efforts are being made to provide links between the two databases.

METHODS

Relational database format

The organization of the database should allow for flexible grouping of data, so that a researcher may, for example, retrieve all experiments that have used a certain technique to study any protein or alternatively retrieve all experiments performed on a particular protein. Protein NMR data (as a reflection of protein conformation and reactivity) are dependent on the solution conditions: pH, temperature, salt concentration, concentrations of co-factors or inhibitors, etc. Because of this, the data must be structured carefully in order to allow, as much as is practical, each experimentally-derived datum to be related back to the conditions under which it was obtained. Of the commercial database management systems (DBMS) currently available to a wide range of platforms, we find that a relational design is the kind best suited to such careful structuring of data. A relational design provides additional benefits in the form of error checking and flexibility of input and retrieval. Owing to the nature of relational databases, the database can be redesigned incrementally to adapt to advances in the field without perturbing the existing data and organization.

A relational database is based on the simple idea of representing all data in tables, also called relations. The data are broken down into elemental parts, and a table is created to store several related aspects (called *attributes*) of entities that the database describes. In a typical business application, the tables might describe customers, equipment parts, and employees; in the protein NMR database, tables describe proteins, experiments, sample conditions, authors of papers, chemical shift assignments, etc. (Fig. 1).

The concept of the *assignment group* (Fig. 2) as the central database entity arose from the common practice of grouping NMR data (chemical shifts, coupling constants, etc.) derived from a set of NMR experiments on a particular molecular species carried out under similar solution conditions. Authors generally are careful in reporting the conditions (pH, temperature, concentrations, etc.) under which data were obtained, but they normally do not specify the particular NMR experiment (2QF COSY, NOESY, etc.) from which a particular datum (chemical shift, coupling constant, etc.) was taken. Where authors have made distinctions in the solution conditions, separate assignment groups are established for each set of conditions. This allows users of the database to combine data obtained under different conditions, if desired, or to preserve the original distinctions. In cases where authors lump together data obtained under a (narrow) range of conditions, the data are put into a single assignment group with ranges specified for variable conditions. Conditions are also given as ranges when authors have provided error estimates on experimental solution variables. For example, if a paper states that '... the chemical shift data were determined at pH 5.0 and 5.5 at protein concentrations between 1.2 and 3 mM', the assignment group would use a pH range of 5.0–5.5 and a protein concentration range of 1.2–3.0 mM. The reporting of ranges rather than points creates computational complexity but provides a fairer estimate of the precision of the data. Another example would be the lumping together of data determined in ^1H_2O and ^2H_2O. In cases where the isotope effect on the chemical shift or other parameters can be ignored (or is neglected by authors in their reporting of the data), the results would be placed into a single assignment group. In cases where the isotope effect cannot be ignored (for example in reporting ^{15}N chemical shifts), separate assignment groups must be used. The preservation in the database of important distinctions such as these hinges on the diligence of authors in reporting them in their publications.

Special tables are created for experiments in which one of the sample conditions is a variable. Such studies include pH titration, variable temperature, and $^1H/^2H$ exchange experiments. For such studies the *assignment group* table would be inappropriate since it specifies constant solution conditions.

In anticipation of an increased need to use information from more than one biological database at a time, we have included links to the Protein Identification Resource, Protein Data Base, Chemical Abstracts Service, and CarbBank databases, via the table *Link_to_other_Database* (Fig. 3).

In the examples of tables shown in Fig. 2, the text names were written out completely for clarity. Actually, database entities, such as *atom names*, *protein names*, and *organism names*, are stored only once, in a lookup table that associates a unique integer with each name (Fig. 4). The unique integer is used in place of the text name in all other places where it is needed. The main purpose of this is to save space, but it also prevents redundancy and the occurrence of multiple spellings

A

table Reference

ref_ID	pub_ID	volume	issue	start page	end page	year	title
250	8 (JBC)	264	32	18907	18911	1989	Determination of the Secondary Structure of Interleukin-8 by Nuclear Magnetic Resonance...
303	20 (Cell)	59	--	573	580	1989	The Structure of the Antennapedia Homeodomain determined by NMR Spectroscopy...
392	8 (JBC)	258	13	8235	8239	1983	Comparative Nuclear Magnetic Resonance Studies of High Potential Iron-Sulfur Proteins
21	3 (EJB)	169	--	201	207	1987	Structural and Dynamical Comparison of alpha, beta, and gamma forms of murine epidermal growth...

B

table Shift

group_ID	residue_numb	aa_name	atom_name	shift_value
40	7	ALA	H^{α}	5.00
40	25	LYS	H^{N}	8.16
40	4	SER	H^{N}	8.00
252	7	PHE	C^{α}	59.00
275	27	SER	H^{α}	4.61
324	32	HIS	$H^{\beta 2}$	3.05

C

table Reporting_Standard

group_ID	nucleus	peak_ID	accuracy
40	^{1}H	1 (methyl-hydrogens of TSP)	0.01
252	^{13}C	3 (dioxane carbons)	0.05

Fig. 1. Examples of *tables* in the protein NMR relational database. (A) The *reference* table that describes journal articles, monographs, and book chapters. Note that the actual name of each journal is displayed in this figure only for clarity. Internally, only the integer pub_ID is stored in this table. This saves storage space, speeds retrieval time, and helps prevent the occurrence of multiple spellings, including misspellings. Any time a user views this information, it is processed so that he sees the publication name, not the integer that corresponds to it. (B) The *shifts* table that stores assignments of NMR chemical shifts to specific atoms. (C) The *reporting standards* table. Each chemical shift assignment is associated with a particular protein under particular experimental conditions (group_ID, see Fig. 2) and with a particular atom of a particular amino acid residue of that protein (aa_name and atom_name). Each row in the *shifts* table (Fig. 1B) identifies one chemical shift assignment. The reporting standard for each chemical shift can be found by matching the *group_ID* in the *shifts* entry (Fig. 1B) to the *group_ID* in the *reporting_standard* table (Fig. 1C). This matching is performed for the user by the relational algebra. The *shifts* table is not stored as a permanent table in the database in this form, but is created dynamically when needed as a *view* (see text).

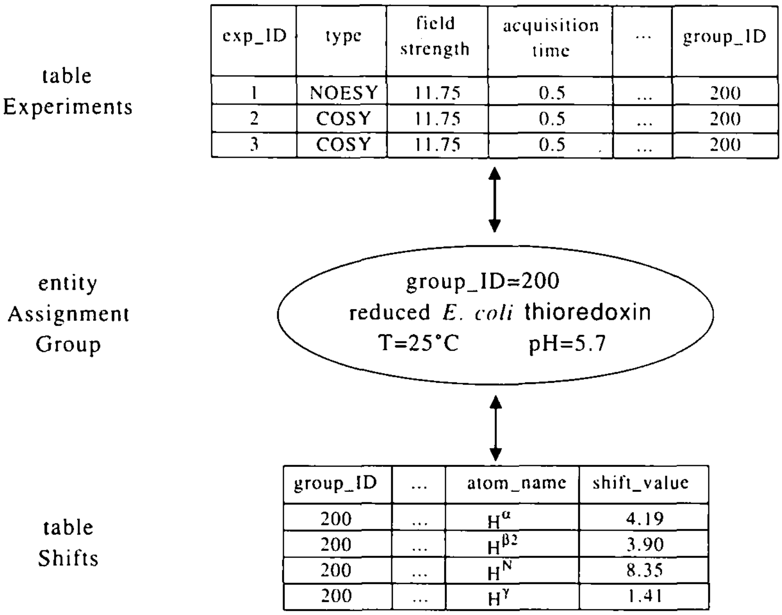


table Experiments

exp_ID	type	field strength	acquisition time	...	group_ID
1	NOESY	11.75	0.5	...	200
2	COSY	11.75	0.5	...	200
3	COSY	11.75	0.5	...	200

entity Assignment Group

group_ID=200
reduced *E. coli* thioredoxin
T=25°C pH=5.7

table Shifts

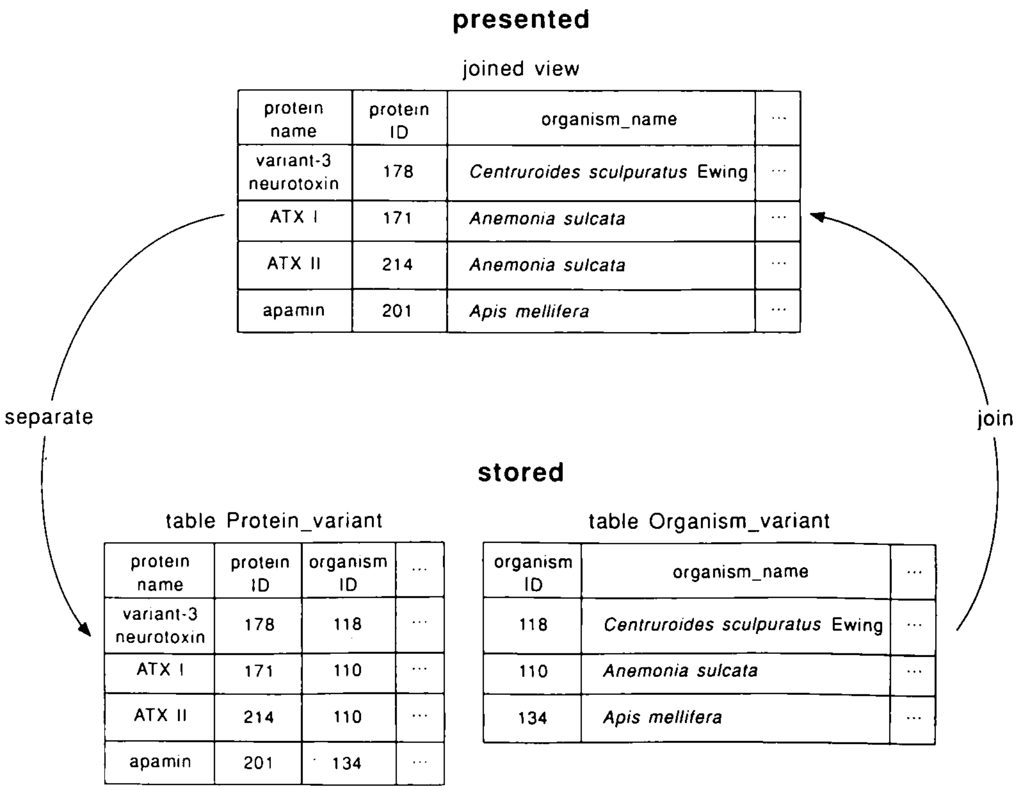
group_ID	...	atom_name	shift_value
200	...	H^{α}	4.19
200	...	$H^{\beta 2}$	3.90
200	...	H^{N}	8.35
200	...	H^{γ}	1.41

Fig. 2. The concept of *assignment group* was invented to realistically model the level of reporting of data and experiments. It links an *experiment* table, which contains data from a set of experiments carried out with a particular protein under similar solution conditions, to a *shifts* table, which holds the chemical shifts and their assignments.

table Link_to_other_Database

protein variant	other database	accession number	expression
164 (*S.aureus* nuclease H124L)	4 (Chem. Abstracts)	9013-53-0	—
164 (*S.aureus* nuclease H124L)	2 (NBRF)	Ncsaf	83-205,L,207-231
164 (*S.aureus* nuclease H124L)	3 (Brookhaven)	2SNS	1-123,L,125-149
164 (*S.aureus* nuclease H124L)	1 (BioMagResBank)	165 (*S.aureus* Nuclease pCQV2)	8-130,L,132-156

Fig. 3. The *Link_to_other_Database* table allows a researcher to relate data in *BioMagResBank* to data in other databases. When another database is found to contain an entry to the same or closely related protein, numbers are inserted into the *Link_to_other_Database* table that identify the other database, the accession number of the (related) protein, and the relationship between that protein and the one in *BioMagResBank* (the notation system we use was shown to us by Drs. David George and Winona Barker). For example, in the second row, we see that the sequence for nuclease H124L in *BioMagResBank* can be derived from the entry stored in the NBRF database as protein 'Ncsaf' (which contains a leader sequence) by concatenating amino acids 83 through 205, inserting a leucine, followed by amino acids 207–231 to obtain a sequence 149 amino acids long. The sequence of nuclease H124L can be obtained from that of another nuclease stored in *BioMagResBank* (*S. aureus* nuclease pCQV2, a recombinant protein with a short amino-terminal extension) by taking the substring amino acids 8 through 130, by then inserting a leucine, and by continuing with amino acids 131 through 156.



joined view

protein name	protein ID	organism_name	...
variant-3 neurotoxin	178	*Centruroides sculpuratus* Ewing	...
ATX I	171	*Anemonia sulcata*	...
ATX II	214	*Anemonia sulcata*	...
apamin	201	*Apis mellifera*	...

table Protein_variant

protein name	protein ID	organism ID	...
variant-3 neurotoxin	178	118	...
ATX I	171	110	...
ATX II	214	110	...
apamin	201	134	...

table Organism_variant

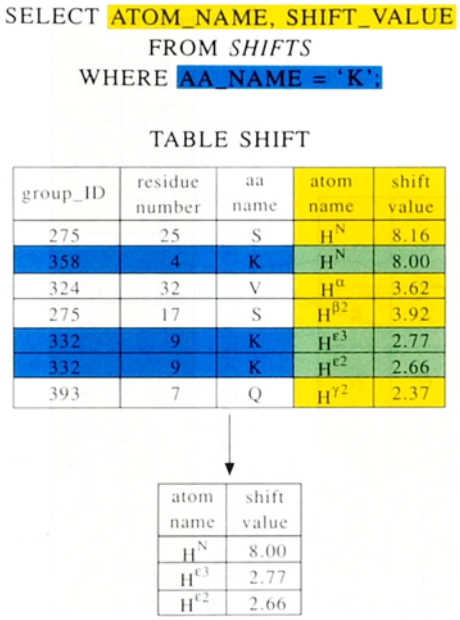
organism ID	organism_name	...
118	*Centruroides sculpuratus* Ewing	...
110	*Anemonia sulcata*	...
134	*Apis mellifera*	...

Fig. 4. Example of a *join*. The *joined view* shown at the top is constructed from two stored tables: *Protein variant* and *Organism variant*. Rather than write the name '*Anemonia sulcata*' for each description of a protein from that organism, what is done internally is to store the integer *110* in the column organism_ID of the table *Protein_variant*. This prevents the occurrence of multiple spellings, which would cause incomplete retrieval of data in some situations. Upon retrieval, the relational algebra *joins* rows from the two tables. The tables shown are fragments of the real tables: other descriptive attributes of proteins and organisms have been omitted here for the sake of clarity.

that occur in other types of databases. The *join* operation of the relational algebra facility of the software (Fig. 4) is used to combine any occurrence of an identifying integer automatically with the text name from the lookup table to create readable text. The user is never required to manipulate the 'bare numbers' directly.

The relational algebra operations (*project*, *union*, *select*, *join*, etc.) can be used to combine data from related tables in a defined fashion (Fig. 5). To invoke the relational algebra, the user writes a *query*. The language in which queries are written resembles symbolic logic. The *query language* used by *BioMagResBank* is *Standard Query Language (SQL)*. Commonly used queries can be 'packaged', so that users can obtain the same results by invoking a simple command rather than by composing an SQL query.

It is the responsibility of the database designer to ensure that the database is in 'normal form'. A normal form is a compact representation of the data (Fig. 6). When the database is in normal form, data are neither created nor lost during the execution of queries.



```
SELECT ATOM_NAME, SHIFT_VALUE
       FROM SHIFTS
     WHERE AA_NAME = 'K';
```

TABLE SHIFT

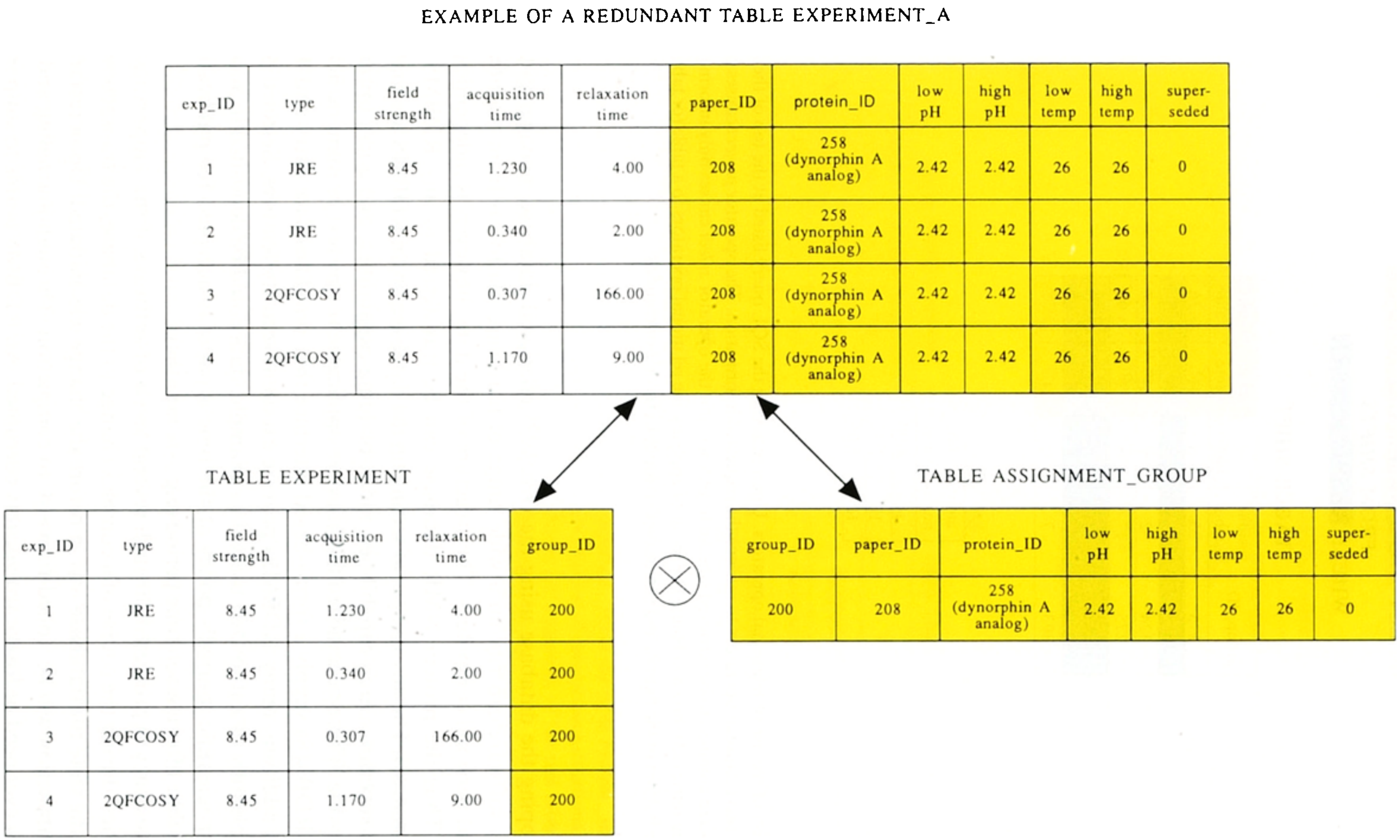
group_ID	residue number	aa name	atom name	shift value
275	25	S	H^N	8.16
358	4	K	H^N	8.00
324	32	V	H^α	3.62
275	17	S	$H^{\beta 2}$	3.92
332	9	K	$H^{\varepsilon 3}$	2.77
332	9	K	$H^{\varepsilon 2}$	2.66
393	7	Q	$H^{\gamma 2}$	2.37

atom name	shift value
H^N	8.00
$H^{\varepsilon 3}$	2.77
$H^{\varepsilon 2}$	2.66

Fig. 5. Demonstration of the use of relational operators. Upon receiving the SQL query listed at the top.of the figure, the SQL parser invokes two relational operators to retrieve the requested information. The operator *project* causes only information from the desired columns to be retrieved (shown in yellow), and the operator *select* causes only information from the desired rows to be retrieved (shown in blue). The result of applying the relational operators to a table (or tables) is itself a table (shown at the bottom of the figure), and can be used as input to further operations.

We are developing the database using the *Oracle*™ relational database management system (RDBMS). This software can run on a wide range of hardware platforms from mainframes to workstations and personal computers. The structure of the data, however, is independent of the RDBMS used. Thus the protein NMR database can be transferred to any RDBMS (public domain or commercial). The advantages of setting up the database on an RDBMS are that we can benefit from the superior search capabilities, index-building capabilities, and query optimization such systems provide. These generalized facilities are applicable to any text or numeric field in the database. Most RDBMSs supply a forms-based interface that simplifies data entry and database development. In addition, many RDBMSs provide excellent security features. Selected data can be made accessible only to authorized users. For example, data could be deposited in accordance with a particular scientific journal's policy at the time a paper is accepted for publication, but released to the public after a given time period (e.g., six months).

Database consistency is protected through several levels of software; if the supporting operating system 'crashes' during the course of a complex change to the database, the transaction will be 'rolled-back' to the state the database was in before the change was initiated. Upon restarting the RDBMS, the transaction will be 'rolled-forward' to completion. Through journaling and regular backups, no data ever need be lost or irrevocably corrupted. An RDBMS is able to support access



EXAMPLE OF A REDUNDANT TABLE EXPERIMENT_A

exp_ID	type	field strength	acquisition time	relaxation time	paper_ID	protein_ID	low pH	high pH	low temp	high temp	super-seded
1	JRE	8.45	1.230	4.00	208	258 (dynorphin A analog)	2.42	2.42	26	26	0
2	JRE	8.45	0.340	2.00	208	258 (dynorphin A analog)	2.42	2.42	26	26	0
3	2QFCOSY	8.45	0.307	166.00	208	258 (dynorphin A analog)	2.42	2.42	26	26	0
4	2QFCOSY	8.45	1.170	9.00	208	258 (dynorphin A analog)	2.42	2.42	26	26	0

TABLE EXPERIMENT

exp_ID	type	field strength	acquisition time	relaxation time	group_ID
1	JRE	8.45	1.230	4.00	200
2	JRE	8.45	0.340	2.00	200
3	2QFCOSY	8.45	0.307	166.00	200
4	2QFCOSY	8.45	1.170	9.00	200

⊗

TABLE ASSIGNMENT_GROUP

group_ID	paper_ID	protein_ID	low pH	high pH	low temp	high temp	super-seded
200	208	258 (dynorphin A analog)	2.42	2.42	26	26	0

Fig. 6. Demonstration of data normalization. Information in the large table at the top of the figure (*Experiment_A*) can be stored more efficiently as two separate tables (*Experiment* and *Assignment_Group*). By using the relation operator *join* on the *group_ID* columns of the two tables (*Experiment* and *Assignment_Group*), we can simulate the table *Experiment_A*.

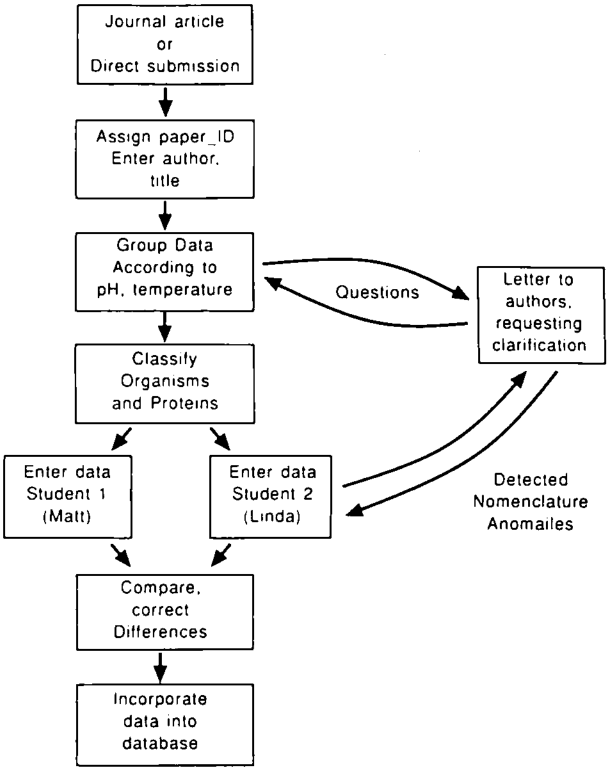


table Track_A

paper ID	shifts 1	shifts 2	shifts_to main	authors	unusual amino_acids	folder
3	Linda	Cate	0 (not done)	1 (done)	0 (none)	compare shifts
12	Linda	0	0 (not done)	1 (done)	1 (has)	for Matt
27	Linda	Matt	1 (done)	1 (done)	0 (none)	NOES

Fig. 7. (Top) Flow chart showing how protein NMR information is handled by the database staff. (Bottom) One of the tables used by *BioMagResBank* staff to track journal articles in the process of being incorporated into the database.

by many users simultaneously, and will present a consistent view to each user. In addition, a relational database can grow incrementally. As soon as any new datum passes certification, it can be incorporated incrementally into the database. Indices will be updated automatically and rebalanced periodically for optimal performance.

Scope of the database

Our present effort is focused on solution NMR studies of proteins or protein fragments. We have limited database entries to studies that meet the following criteria: the molecule investigated must be a protein or peptide of 12 residues or more; the NMR data must have been acquired in solution; the NMR data must be assigned to one or more specific atoms in a designated residue in a known sequence; and the data must have appeared in the primary refereed literature. The new literature is scanned routinely to identify potential new entries; the older literature has been entered in a less systematic way by relying on review articles and reprint collections for relevant entries. The database format we have developed is general and can accommodate amino acids and smaller peptides or other classes of biomolecules.

Data entry

Several tables in the database itself are used to track the progress of journal articles being considered for inclusion in the database (Fig. 7). Given the complexity of NMR data and the variability of reporting practices, each journal article requires detailed scrutiny by an annotator prior to data entry. Each protein entry in the *BioMagResBank* database, whenever possible, is matched against *PIR*, *PDB*, and Chemical Abstract Services (*CAS*) entries. The organism from which the protein was derived is classified by genus, species, and variant, and genus and species are checked for occurrence in the *PIR* and *CAS* databases. If the species is not found, then a taxonomist is consulted to verify that the genus and species names are currently accepted. Proteins are classified according to IUB convention (IUB, 1984) if they have an enzymatic activity. The amino acid sequence presented in the paper is checked against the sequences of proteins with the same name (or similar names) that occur in other databases: *PIR*, *PDB*, or *CAS*. Questions about experimental protocols and reporting standards are referred to spectroscopists. Data entry itself is not a trivial operation because of variations in nomenclature such as residue number specification or atom specification. Most time consuming is the careful reading of the entire paper, particularly the materials and methods section, needed in order to establish experimental conditions including temperature, pH, and subtle aspects of the protein species under consideration such as chemical modification, presence of a bound ligand, or co-factor.

We have taken advantage of properties of the relational database format in order to implement automatic error checking during the data-entry process. As an example, by using our automated data entry program an undergraduate data-entry worker uncovered over twenty instances in the published literature of chemical shift assignments to non-existent atoms in protein structures. The entry program acts as a filter and rejects impossible entries (e.g., an assignment to an alanine $^1H^\gamma$). By contrast, the creation of a similar level of error checking in a free-text database would require elaborate programming. As a check against errors introduced during data entry, we have each datum entered by two different persons with the preliminary results stored in two temporary relations. We then use the algebraic query language, *SQL*, to ask for and flag any discrepancies so that they can be checked and reconciled before the data set is made a permanent part of the database (Fig. 8).

Database distribution

We propose to distribute the database in three different formats to match the computer equipment and expertise of users. The simplest format is a *flat file*, in which data would be available as text in a fixed format. Keywords would be used to identify each type of data. This format, which can be used by any laboratory, loses much of the power obtained by storing the data in a relational format. In *carrier file* format, the data would be separated from the *Oracle* RDBMS that has been used in developing *BioMagResBank*. The data can then be transferred to any relational data-

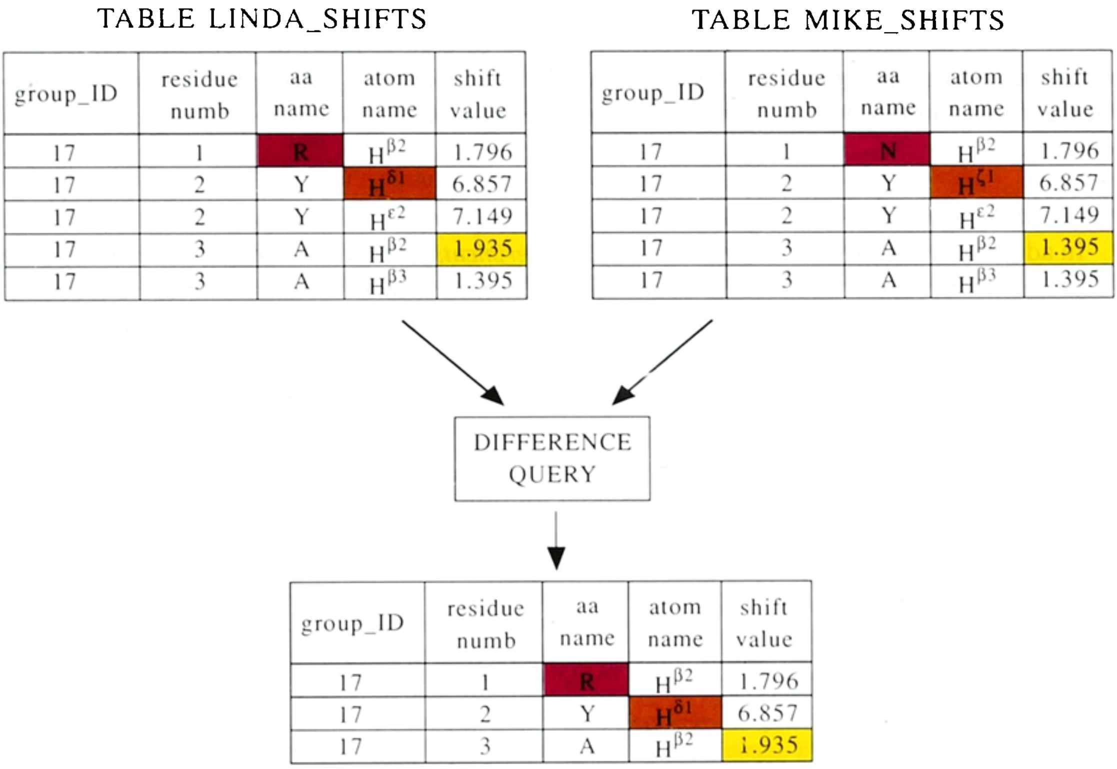


TABLE LINDA_SHIFTS

group_ID	residue numb	aa name	atom name	shift value
17	1	R	$H^{\beta2}$	1.796
17	2	Y	$H^{\delta1}$	6.857
17	2	Y	$H^{\varepsilon2}$	7.149
17	3	A	$H^{\beta2}$	1.935
17	3	A	$H^{\beta3}$	1.395

TABLE MIKE_SHIFTS

group_ID	residue numb	aa name	atom name	shift value
17	1	N	$H^{\beta2}$	1.796
17	2	Y	$H^{\zeta1}$	6.857
17	2	Y	$H^{\varepsilon2}$	7.149
17	3	A	$H^{\beta2}$	1.395
17	3	A	$H^{\beta3}$	1.395

DIFFERENCE QUERY

group_ID	residue numb	aa name	atom name	shift value
17	1	R	$H^{\beta2}$	1.796
17	2	Y	$H^{\delta1}$	6.857
17	3	A	$H^{\beta2}$	1.935

Fig. 8. Example of the routine used to screen out errors introduced during data input. Each data set is entered temporarily into separate relations by two different workers (Linda and Mike in the example shown). Two SQL difference queries are performed in order to uncover any discrepancies. Here the query:

```
   select group_ID, residue_numb, aa_name, atom_name, shift_value
   from linda_shifts
   where group_ID = 17
minus
   select group_ID, residue_numb, aa_name, atom_name, shift_value
   from mike_shifts
   where group_ID = 17
```

finds rows that are in the table *linda_shifts*, but not in *mike_shifts*. Three types of errors were detected here: (1) In the first row, Mike correctly abbreviated arginine as *R*, but Mike incorrectly used *N*. (2) In the second row, one of the workers used the wrong atom name. (3) In the third retrieved row, Mike entered 1.395 as the shift value and Linda entered 1.935. A second query that subtracts *linda_shifts* from *mike_shifts* will detect any remaining differences. After differences are detected, checked, and corrected, the shifts are made part of the permanent database. (All tables in this figure are views created by joining the atom lookup table with the actual internal representation of shifts.)

base management system in order to accommodate groups that use an RDBMS other than *Oracle*. This would preserve the power of indexing and the relational algebra. SQL queries would not be affected. Programs that interact with *BioMagResBank* by means of a third generation computer language (e.g. *FORTRAN* or *C*) with system calls to the database might need slight modification to accommodate system dependencies. We also plan to provide a *network server* to *BioMagResBank*. This would provide automatic responses to SQL queries or 'canned' commands.

Database personnel could supply responses to ad hoc natural language queries. We expect that a major use of the database will be to retrieve data to be used as input to preexisting NMR and structural analysis programs. Lookup tables can be used in order to convert the data into the particular format (atom nomenclature, coordinate system, etc.) expected by the analysis program (Fig. 9).

Data catalogued

In addition to entities such as *proteins*, *NMR experiments*, *organisms*, and *shift assignments*, the database can now describe *coupling constants*, NOE measurements (*distance ranges*, *NOE buildup rates*), and pK_a *values* (Fig. 10). We have decided to wait until more of a consensus has developed concerning protocols before entering data on protein structures derived from NMR data.

RESULTS AND DISCUSSION

In the past few months the database has been transformed, from a small prototype created to test the design, into a comprehensive, organized collection. The current database contains information from over 1100 papers. More than 2300 different authors, 1500 assignment groups, and

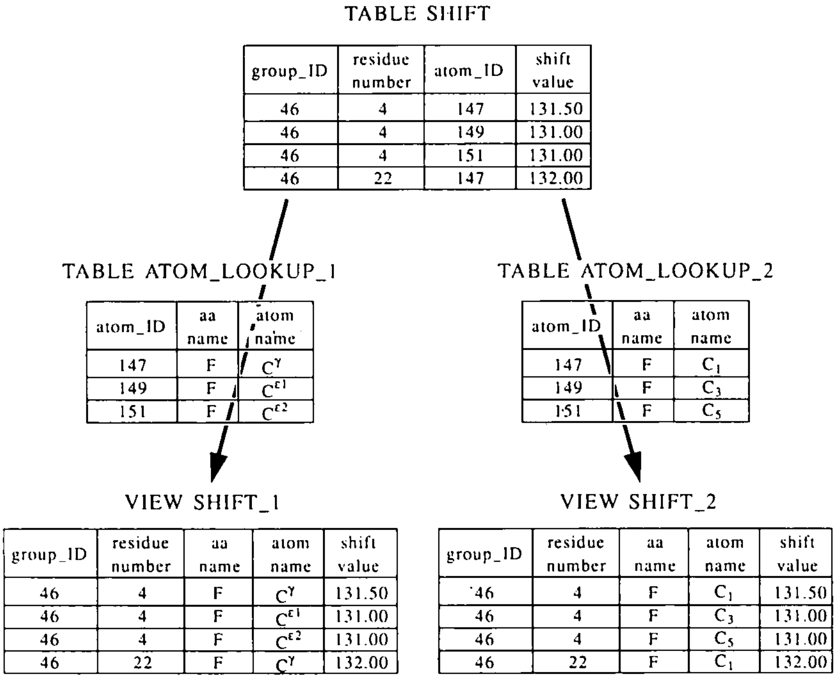


TABLE SHIFT

group_ID	residue number	atom_ID	shift value
46	4	147	131.50
46	4	149	131.00
46	4	151	131.00
46	22	147	132.00

TABLE ATOM_LOOKUP_1

atom_ID	aa name	atom name
147	F	C^{γ}
149	F	$C^{\epsilon1}$
151	F	$C^{\epsilon2}$

TABLE ATOM_LOOKUP_2

atom_ID	aa name	atom name
147	F	C_1
149	F	C_3
151	F	C_5

VIEW SHIFT_1

group_ID	residue number	aa name	atom name	shift value
46	4	F	C^{γ}	131.50
46	4	F	$C^{\epsilon1}$	131.00
46	4	F	$C^{\epsilon2}$	131.00
46	22	F	C^{γ}	132.00

VIEW SHIFT_2

group_ID	residue number	aa name	atom name	shift value
46	4	F	C_1	131.50
46	4	F	C_3	131.00
46	4	F	C_5	131.00
46	22	F	C_1	132.00

Fig. 9. Example of the generation of tables of data that employ different nomenclature systems from stored tables and lookup (conversion) tables. The two tables at the bottom are actually *views*, i.e. the tables are not stored in the computer but are created dynamically from the underlying tables whenever needed. This translation can be evoked from a *report*, which would also format the data text spatially as input as required by the target program. Translation from one-letter amino acid abbreviations to three-letter abbreviations or full names can be accomplished in the same fashion.

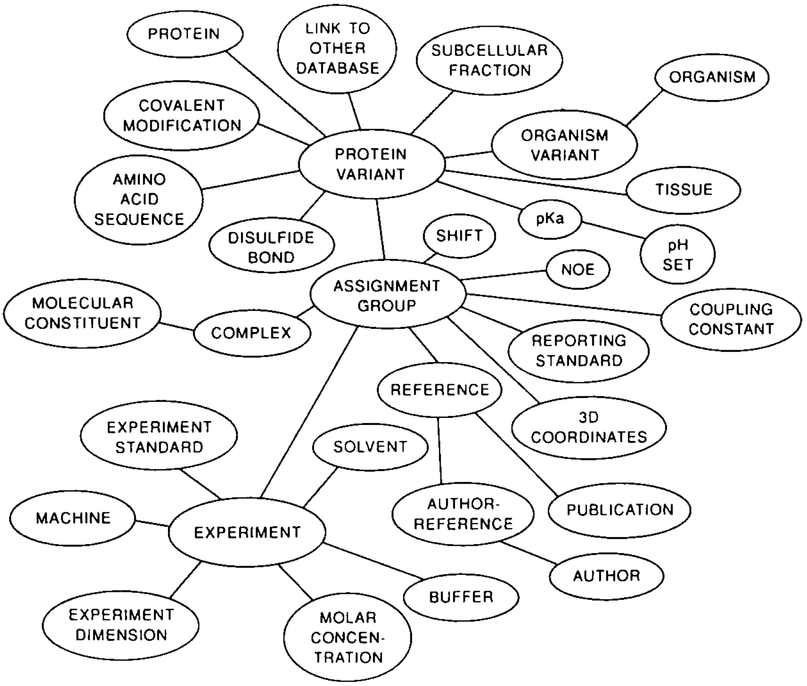


Fig. 10. Entities and relationships modeled in *BioMagResBank*. Each oval corresponds to a table (entity) in the database (not all the attributes of each table can be displayed in this figure). A line connecting two entities indicates that there is some natural relationship between the two entities. For example, every *reference* (journal article or book chapter) has at least one *author* and appears in a *publication* (journal, monograph, or proceedings). Every shift assignment or coupling constant can be related back (by means of the entity *assignment group*) to a particular protein.

20000 shifts have been included*. Information can be retrieved as indexed by authors, protein, organism, sequence, amino acid, pH of data collection, particular NMR experiment, or any other characteristic described in the database.

Our guiding principles have been to create a design directly accessible to scientists at various levels of computer proficiency and to support automatic manipulation of and reasoning about the data by programs written in higher level languages. The desire for a clean, simple design has been tempered by a realistic acceptance of the lack of a standard reporting format and of the limits of experimental accuracy.

*Compilation of data has been hindered by incomplete reporting in journal articles. The most common problems have been: the failure to define experimental or reporting chemical shift standards, incorrect naming of atoms, undefined experimental conditions such as temperature or pH, and lack of clarity in reporting spectral parameters. Over 200 letters have been sent to authors requesting clarification or additional details. Responses to these letters (at a level of about 50%) have allowed us to release assignment groups that were placed on hold.

ACKNOWLEDGEMENTS

We are indebted to members of the CODATA Task Group on Biological Macromolecules for their advice and encouragement in developing *BioMagResBank*, in particular, Professor A. Tsugita (Science University of Tokyo), Drs. W.C. Barker and D.G. George (Protein Identification Resource). Taxonomic advice was kindly provided by Dr. Andrzej Elzanowski (Martinsrieder Institut für Proteinsequenzen).

This research was supported by grants LM04958 from the National Library of Medicine, National Institutes of Health, RR 02301 from the Biomedical Research Technology Program, National Center for Research Resources, National Institutes of Health, and #1104 from the New Energy and Industrial Development Organization of Japan.

REFERENCES

Division of Biology and Medicine, International Society of Magnetic Resonance (1990) *Guidelines for the Publication of Protein and Nucleic Acid Structures Derived from NMR Data* (Task force members: V.F. Bystrov, C.M. Dobson, G.C.K. Roberts, G. Wagner, and K. Wüthrich).

Fesik, S.W., Gampe, R.T. and Zuiderweg, E.R.P. (1989) *J. Am. Chem. Soc.*, **111**, 770-772.

Gao, Y., Boyd, J., Williams, R.J.P. and Pielak, G.J. (1990) *Biochemistry*, **29**, 6994-7003.

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

IUB (1984) *Enzyme Nomenclature (Recommendations (1984) of the Nomenclature Committee of the International Union of Biochemistry)*, Academic Press, New York, NY, 656 pp.

Jardetzky, O. (1989) *Science*, **246**, 431.

Kay, L.E., Clore, G.M., Bax, A. and Gronenborn, A.M. (1990) *Science*, **249**, 411-414.

Markley, J.L. and Ulrich, E.L. (1984) *Annu. Rev. Biophys. Bioeng.*, **13**, 493-521.

Markley, J. (1990) *Methods enzymol.*, **176**, 12-64.

Markley, J.L., Seavey, B.R., Alexandrescu, A.T., Darba, P., Hinck, A.P., Loh, S.N., McNemar, C.W., Mooberry, E.S., Oh, B.-H., Stockman, B.J., Wang, J., Westler, W.M., Zehfus, M.H. and Zolnai, Zs. (1990) in *Protein Engineering. Protein Design in Basic Research, Medicine and Industry* (Eds, Ikehara, M., Oshima, T. and Titani, K.), Springer-Verlag, New York, NY, pp. 285-290.

Stockman, B.J. and Markley, J.L. (1990) *Advances in Biophysical Chemistry, Vol. 1* (Ed. Bush, C.A.), JAI Press Inc., Greenwich, CT, pp. 1-45.

Ulrich, E.L., Markley, J.L. and Kyogoku, Y. (1989) *Prot. Seq. Data Anal.*, **2**, 23-37.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley-Interscience, New York, NY, 292 pp.

Zuiderweg, E.R.P. and Fesik, S.W. (1989) *Biochemistry*, **28**, 2387-2391.

APPENDIX

Check list for protein NMR data to be submitted for publication or for database inclusion

John L. Markley, Elizabeth A. Farr and Beverly R. Seavey

Biochemistry Department, College of Agricultural and Life Sciences, University of Wisconsin-Madison, 420 Henry Mall, Madison, WI 53706-1569, U.S.A.

The following check list has been devised to assist authors in preparing protein NMR data for publication. The list is more general and more comprehensive than the set of recommendations recently adopted by the Division of Biology and Medicine, ISMAR (1990). The protein NMR field is evolving rapidly, and it is to be expected that a list of this kind will change over time. Some of the sections (e.g., Section 1 on protein description) are pertinent to all sets of experimental protein NMR data; other sections concern only certain classes of experiments. BioMagResBank (Ulrich et al., 1989; Seavey et al., 1991) collects data from published protein NMR articles and from supplementary material deposited with publications. In the case of journals that do not accept supplementary materials, these can be submitted directly to BioMagResBank, and a note to this effect can be placed in the article. BioMagResBank does not accept data that are not associated with a refereed publication. Authors of papers reporting protein NMR data are encouraged to contact BioMagResBank in advance of the submission of their manuscript to receive a reference identification number. This number will be issued upon receipt of the manuscript's title, list of authors, and name of the journal to which the paper will be submitted. The title page of the paper should contain a footnote: 'protein NMR data will be deposited in BioMagResBank under ref_ID ####', where #### is the reference identification number supplied.

1. Protein description (to be reported in the abstract as well as in the introduction or methods section of the publication).
   1.1. Protein name (should be reported in the abstract as well as in the introduction or methods section of the publication)[a].
      1.1.1. Trivial name.
      1.1.2. ECE (IUB, 1984), PIR (Protein Identification Resource), or CAS (Chemical Abstracts Service) designation if known.

[a] Include any common name of the protein and an accepted classification if it exists. Electron-transfer proteins, such as cytochromes, are not assigned ECE numbers, but an appendix in the IUB (1984) book on enzyme nomenclature describes the classification of these proteins. The IUPAC-IUB Commission (1975a) has discussed the nomenclature of iron-sulfur proteins.

1.2. Protein source (to be reported in the abstract as well as in the introduction or methods section of the publication).

1.2.1. If biological: the origin of the protein (genus, species; tissue, cell type, organelle, etc., if pertinent).

1.2.2. If recombinant: origin of the gene; species in which it was expressed[b].

1.2.3. If synthetic: synthetic procedure.

1.2.4. If semisynthetic: origin of components and method of synthesis.

1.3. Protein sequence and covalent structure[c].

1.3.1. Provide complete sequence of the protein molecule studied, or a reference to that sequence. (If the molecule is a fragment of a larger protein or a recombinant protein with an extension or deletion, this should be noted explicitly.)

1.3.2. Indicate known disulfide bridges, unusual or modified amino acids, and co-factors (covalent or non-covalent)[d].

1.4. Protein isotopic composition.

1.4.1. Natural abundance.

1.4.2. Isotopically enriched[e].

1.4.2.1. Isotopic species.

1.4.2.2. Location, enrichment levels, and pattern of enrichment.

[b] Identify the host strain of organism that produced the protein, the plasmid if applicable, and for recombinant proteins the identity of the original source (including the strain, if known) of the gene (e.g., organism from which it was cloned, synthetic gene). For example: 'staphylococcal nuclease H124L was produced by a modified version of the plasmid pTSN1, grown in the host *E. coli* strain BL21-(DE3); the original clone was from the Foggi strain of *Staphylococcus aureus* but was mutated to yield leucine at position 124; the protein produced by the expression system has a sequence identical to the nuclease isolated from the V8 strain of *S. aureus*.

[c] In order to simplify storage and computation, BioMagResBank stores all peptide sequences as starting with residue number 1 and assigns sequential integers to adjacent residues. When known, the entire amino acid sequence should be given, even if spectral assignments are not reported for certain residues (e.g., an amino terminal stretch). Numbering by homologous residues, although convenient for some comparison purposes, would introduce an extra level of complexity into the database. Protein fragments are considered as separate molecules, and are numbered starting with their actual amino terminus as residue 1. Fragments of larger proteins should be carefully described as such at the first point in a publication at which they are named and described. An IUPAC-IUB (1972) publication discusses nomenclature for synthetic polypeptides. The IUPAC-IUB Commission (1975b) has presented nomenclature for peptide hormones.

[d] The IUPAC-IUB (1984) standards for naming amino acids provide the names for all standard and many modified or unusual amino acids. Additional examples are given by Cohn (1984). In cases not covered by this list, BioMagResBank will adhere to the nomenclature adopted by the Chemical Abstracts Service. When authors are not certain of the nomenclature or when a modification is not covered by published nomenclature, it would be best for authors to include a stereospecific representation of the modification in the publication with the pertinent atoms clearly identified. The nomenclature of porphyrins and related compounds has been described by Bonnett (1978).

[e] In cases of isotopic labeling, the means of labeling should be described. The isotope, position(s) labeled, and level of incorporated isotope should be specified. For example, '.. tyrosine with 95% ^{13}C at the δ1 position was prepared by chemical synthesis; the label was introduced into the protein by feeding the amino acid to the organism that produced the protein; the overall level of enrichment at this site in the protein was estimated to be 80%'.

2. Protein solution and experimental conditions[f].
   2.1. Concentration of each protein present (units: molar). (Is the protein present as a monomer, dimer, or higher aggregate at this concentration?)
   2.2. Identity and concentration (isotopic composition) of ligand (inhibitor, co-factor, substrate, protein target, etc.) (units: molar).
   2.2. Salt(s) (provide counter ions, not just '0.5 mM Ca^{2+}'); buffer(s) (units: molar).
   2.3. Solvent composition (including isotopic composition or range of compositions) (units: vol/vol %)[g].
   2.4. pH (value or range, accuracy of the measurement, how measured, any correction for isotope effects).
   2.5. Temperature (value or range, accuracy of the measurement, how measured).
   2.6. Oxidation state of the protein and the means by which the oxidation potential of the solution was adjusted. (For example, 'the protein was reduced by bubbling N_2 gas through the solution in the NMR tube; 1 mg of dithionite was then added, and the NMR tube was sealed'.)
   2.7. Was the protein in its native state under the conditions of the experiment? If non-native, what was the denaturant?
   2.8. Was the protein studied as a complex under these experimental conditions? If so, what is the stoichiometry of the complex studied?

3. NMR experiment (to be described in the experimental section of the publication).
   3.1. NMR hardware: type of spectrometer with any special modifications; probe; special filters; lock signal (if used); spinning rate; temperature controller.
   3.2. Diagram of the pulse sequence used with phase cycling, or a literature reference to the pulse sequence used.
      3.1.1. Description of the experiment: dimensionality, connectivities, coherences.
      3.1.2. Frequencies of nuclei excited and detected.
      3.1.2. List of delays, mixing times, etc., used.
   3.3. Chemical shift reference for each nucleus reported[h].
      3.2.1. Experimental standard reference for chemical shift determination

[f] It is useful to know as much as possible about the composition of the sample and the experimental conditions used in obtaining the NMR data. If any of the data (e.g., chemical shifts) were determined under distinct and different conditions, including solvent isotopic composition (e.g., 10% ^2H_2O and 90% ^1H_2O), this should be noted explicitly. Certain values may be reported as ranges if this more accurately reflects experimental error or procedures. Such information should appear in the headings or footnotes of any table of data (not just in figure captions). For example: '^1H shifts of *Lactobacillis casei* dihydrofolate reductase (1 mM–3mM) at 30°C, pH 6.5 ± 0.5 (400 mM KCl, 50 mM potassium phosphate buffer, 1 mM–3 mM methotrexate reported in 90% H_2O/10% ^2H_2O from internal (0.3 mM) DSS'.

[g] In view of the evidence for solvent $^1H_2O/^2H_2O$ isotope effects on ^{15}N and ^{13}C chemical shifts, it is important that the solvent composition for each experiment be specified clearly when such data are presented.

[h] A typical statement would be: 'proton chemical shifts are reported relative to the methyl resonance of DSS assigned the value of 0.000 ppm; they were actually measured relative to internal dioxane (0.3 mM) which was assumed to have a chemical shift value of 3.76 ppm relative to DSS; shifts are reported to ±0.002 ppm'. Such statements should appear in the experimental section of the publication.

(compound, isotopic composition, nucleus referenced, which signal from that nucleus; concentration; internal or external).

3.2.2. Reporting standard – if different from the experimental standard (compound, nucleus referenced, isotopic composition; concentration; conversion factor from the experimental standard and the reporting standard; internal or external).

4. Conventions used for naming parts of proteins and associated molecules.

4.1. Residue numbering: in the database, amino residues will be numbered sequentially from the actual amino terminus (numbering systems that refer to homology, consensus sequences, full-length proteins, etc., will be ignored).

4.2. Atom designations within residues: the IUPAC-IUB (1970) numbering system will be used[i].

4.2.1. Prochiral assignments: the IUPAC-IUB nomenclature has provision for unambiguous designation of prochiral atoms or groups (e.g., $H^{\beta 2}$ and $H^{\beta 3}$, Val $(H^{\gamma 1})_3$ and $(H^{\gamma 2})_3$).

4.2.2. In cases of prochiral ambiguity, the IUPAC-IUB nomenclature *should not be used*, and prochiral spectra assignments should be distinguished by primes or double primes: for example, β, β′.

4.3. Naming of amino acids (including unusual and modified residues): IUPAC-IUB nomenclature should be adhered to as far as possible[d].

4.4. Co-factor labeling[d].

5. Reporting of NMR data.

5.1. Primary data

5.1.1. Raw spectral data. (These are the initial results of an NMR experiment. These data should be saved in the laboratory in which they were collected. Eventually, means may be found for incorporating such valuable data into a database.)

5.1.2. Processed data. (Examples of the processed data should be included in publications to illustrate methods and the validity of the conclusions of the study.)

5.1.3. Parameters derived from NMR spectra.

5.1.3.1. Chemical shifts (indexed to individual assigned nuclei) (units: ppm (parts per million) from a specified reference (see section 3.2).

5.1.3.2. Coupling constants (indexed to assigned pairs of nuclei) (units: Hz).

[i] BioMagResBank has adopted the Greek letter (or Roman equivalent) designation of atoms in residues described by the IUPAC-IUB (1970) commission. This notation is that used by the Protein Data Bank and is generally accepted by the X-ray crystallographic community. This labeling convention gets around ambiguities that exist in other numbering systems, particularly for the rings of aromatic amino acids. The use of *ortho* and *meta* for atoms in tyrosine and phenylalanine is particularly confusing, since different authors use the two possible meanings of these terms. Whenever authors use notation that deviates from the IUPAC-IUB standard, they must provide sufficient description to avoid confusion.

5.1.3.3. NOE or ROE values (or build-up rates) indexed to assigned pairs of nuclei (indicate whether these are from measured peak heights, contour levels, or integrals, and how the measurements were made)[j].

5.1.3.4. T_1, T_2, or $T_{1\rho}$ values (indexed to individual assigned nuclei) (units: seconds)[k].

5.2. Derived data (assigned to a particular atom or group, under specified experimental conditions).

5.2.1. Hydrogen exchange rates (these may be phenomonological rates measured at a particular pH value or the rates fitted to the equation that takes acid and base catalysis of exchange into account).

5.2.2. Protection factors for hydrogen exchange.

5.2.2. Correlation times.

5.2.3. Order parameters.

5.2.4. p*K* values.

5.2.5. Equilibrium constants.

5.2.5.1. Protein conformational equilibria.

5.2.5.2. Ligand binding equilibria.

5.2.6. Dihedral angles and rotamer populations.

5.3. Structural models.

5.3.1. Input parameters: complete list of constraints used in generating the structure and the conditions under which they were obtained (pH, temperature, protein concentration, salt concentration, etc. (see section 2 above)).

5.3.1.1. NOE and/or ROE distance constraints: list of upper and lower distance limits (values for lower limits should be supplied even if they are simply assumed to be van der Waals contacts)[l].

5.3.1.2. Constraints from coupling data[m].

5.3.1.3. Constraints from inferred hydrogen bonding.

5.3.1.4. Constraints from known or assumed covalent structure (disulfides, metal ligation, or other cross links).

5.3.1.5. Explicit functional forms of potentials used as constraints in the structure determination.

5.3.1.6. Other information or assumptions (e.g., information from crystal structures, structural, or NMR databases).

5.3.2. Deduced secondary structure.

[j] For NOEs or ROEs, a range of values from 1 (weakest) to 10 (strongest), for example, may be more informative than the designations 'strong', 'medium', and 'weak'.

[k] Alternatively, relaxation rates (units: s^{-1}) may be given.

[l] Include details on how distances were calibrated.

[m] The IUPAC-IUB (1970) commission has published conventions for specifying dihedral angles and peptide conformation.

5.3.3. Sets of 3D coordinates consistent with NMR results (discuss or provide a reference to the algorithm used for structure calculation and the methods used for determining the agreement between the calculated structure and the NMR data).

5.3.3.1. Protein coordinates with estimates of their accuracy.

5.3.3.2. Bound ligand positions.

5.3.3.3. Bound water positions.

5.3.3.4. Estimates of the quality of the structures: deviations from ideal covalent geometry, numbers of bad contacts both for bond lengths and bond angles (with a reference to the standard covalent geometry used), Ramachandran (φ,ψ) plots.

REFERENCES

Bonnett, R. (1978) In *The Porphyrins, Vol. 1, Part A.* (Ed, Dolphin, D.) Academic Press, New York, NY, pp. 1-27.

Cohn, W.E. (1984) *Methods Enzymol.*, **106**, 3-17.

Division of Biology and Medicine, International Society of Magnetic Resonance (1990) *Guidelines for the Publication of Protein and Nucleic Acid Structures Derived from NMR Data* (Task force members: V.F. Bystrov, C.M. Dobson, G.C.K. Roberts, G. Wagner, and K. Wüthrich).

IUB (1984) *Enzyme Nomenclature (Recommendations (1984) of the Nomenclature Committee of the International Union of Biochemistry)*, Academic Press, New York, NY, 656 pp.

IUPAC-IUB Commission on Biochemical Nomenclature (1970) *Biochemistry*, **9**, 3471-3479.

IUPAC-IUB Commission on Biochemical Nomenclature (1972) *Biochemistry*, **11**, 942-944.

IUPAC-IUB Commission on Biochemical Nomenclature (CBN) (1975a) *Eur. J. Biochem.*, **35**, 1-2.

IUPAC-IUB Commission on Biochemical Nomenclature (1975b) *Biochemistry*, **14**, 2559-2560.

IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN) (1984) *Eur. J. Biochem.*, **138**, 9-37.

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

Ulrich, E.L., Markley, J.L. and Kyogoku, Y. (1989) *Prot. Seq. Data Anal.*, **2**, 23-37.