J Biomol NMR (2012) 54:169–179
DOI 10.1007/s10858-012-9660-3

ARTICLE

PACSY, a relational database management system for protein structure and chemical shift analysis

Woonghee Lee · Wookyung Yu · Suhkmann Kim · Iksoo Chang · Weontae Lee · John L. Markley

Received: 11 April 2012 / Accepted: 8 August 2012 / Published online: 19 August 2012

Abstract PACSY (Protein structure And Chemical Shift NMR spectroscopY) is a relational database management system that integrates information from the Protein Data Bank, the Biological Magnetic Resonance Data Bank, and the Structural Classification of Proteins database. PACSY provides three-dimensional coordinates and chemical shifts of atoms along with derived information such as torsion angles, solvent accessible surface areas, and hydrophobicity scales. PACSY consists of six relational table types linked to one another for coherence by key identification numbers. Database queries are enabled by advanced search functions supported by an RDBMS server such as MySQL or PostgreSQL. PACSY enables users to search for combinations of information from different database sources in support of their research. Two software packages, PACSY Maker for database creation and PACSY Analyzer for database analysis, are available from http://pacsy.nmrfam.wisc.edu.

Keywords PACSY · NMR · PDB · BMRB · SCOP · Database · Structural biology · Bioinformatics

Electronic supplementary material The online version of this article (doi:10.1007/s10858-012-9660-3) contains supplementary material, which is available to authorized users.

W. Lee (✉) · J. L. Markley (✉)
National Magnetic Resonance Facility at Madison, and Biochemistry Department, University of Wisconsin-Madison, Madison, WI 53706, USA
e-mail: whlee@nmrfam.wisc.edu

J. L. Markley
e-mail: markley@nmrfam.wisc.edu

W. Lee · W. Lee (✉)
Structural Biochemistry and Molecular Biophysics Laboratory, Department of Biochemistry, Yonsei University, Seoul 120-749, Korea
e-mail: wlee@spin.yonsei.ac.kr

W. Yu · I. Chang
Department of Physics, Center for Proteome Biophysics, Pusan National University, Busan 609-735, Korea

S. Kim
Department of Chemistry and Chemistry Institute for Functional Materials, Pusan National University, Busan 609-735, Korea

Introduction

The importance of three-dimensional structures of proteins derives from their relevance to biological function. It was recognized early on, when only a handful of X-ray structures of proteins had been solved, that it would be valuable to make the information available from a publicly accessible data bank, and this led to the establishment of Protein Data Bank (PDB) (Bernstein et al. 1977). The data format of the PDB has been extended, and the current Worldwide Protein Data Bank (wwPDB) now encompasses structural data from NMR spectroscopy as well as X-ray crystallography (Berman et al. 2007). Comparisons of three-dimensional structures provide information on evolutionary relationships, and analyses of this kind are available from the SCOP database (Murzin et al. 1995) and the CATH database (a hierarchic classification of protein domain structures, Orengo et al. 1997). Currently most of the structures in the PDB have been solved by either X-ray crystallography (87.7 %) or NMR spectroscopy (11.8 %); but a growing number of structures are being determined by electron microscopy (0.5 %). Although structure determination by NMR spectroscopy has limitations in that it is not as highly automated as X-ray crystallography and not as successful with large proteins or protein complexes,

it offers several interesting features. NMR structures can be solved in solution under molecular conditions similar to those in vivo. NMR can be used to determine dynamic properties of proteins (both local and global). In addition, NMR as a spectroscopic approach can be used to determine thermodynamic and kinetic properties of proteins and their interactions with other molecules. The Biological Magnetic Resonance Bank (BMRB) (Ulrich et al. 2008) provides an archive for the full range of biomolecular NMR data, in addition to its role as the repository of chemical shifts and restraints associated with three-dimensional NMR structures as a partner in the wwPDB (Markley et al. 2008).

Structure calculations from NMR data typically depend on determining a variety of constraints, including distance constraints from NOE measurements, dihedral angle restraints from chemical shifts or spin–spin couplings, and/or projection angles between bond vectors from residual dipolar coupling measurements. It has long been recognized that NMR chemical shifts contain information on local structure, and this is the basis for the approaches used to determine secondary structure from NMR chemical shifts (Eghbalnia et al. 2005b; Wishart and Sykes 1994; Wishart et al. 1992)). Currently, TALOS (Torsion Angle Likelihood Obtained from Shifts and sequence similarity) (Cornilescu et al. 1999) and its successor TALOS+ (Shen et al. 2009a) are the most popular software packages used to predict dihedral angles from NMR chemical shifts for use as angle constraints in structure calculations. The accuracy of such predictions can be improved by making use of homology modeling (Berjanskii et al. 2006). Several software packages have been developed that provide robust determinations of 3D structures from the available constraints: these include CYANA (Güntert 2004), ARIA (Bardiaux et al. 2012), CNS (Brunger et al. 1998), and Xplor-NIH (Schwieters et al. 2003).

Newer approaches to protein NMR data collection and analysis are streamlining and automating the steps in protein structure determination. Reduced dimensionality and sparse sampling approaches (Bahrami et al. 2012; Eghbalnia et al. 2005a; Gledhill and Wand 2012; Hiller et al. 2005; Hyberts et al. 2010; Kim and Szyperski 2003; Schulte-Herbruggen et al. 1999; Stanek and Kozminski 2010) are speeding up NMR data collection. Furthermore, the use of protein modeling approaches along with a chemical shifts as constraints appears very promising, particularly for small proteins (Sgourakis et al. 2011; Shen et al. 2008; Shen et al. 2009b).

Although clear relationships have been found between 3D structure and NMR parameters (e.g., chemical shifts, J-coupling constants, RDC values), tools are lacking that enable the combined analysis of data from the PDB, BMRB, and SCOP databases. One of the reasons for this is that PDB and BMRB data are stored in flat-file formats, versions of the Self-defining Text Archive and Retrieval (STAR) file format (Hall and Spadaccini 1994). As an aid to easier and faster handling of the huge information content of these databases, we have developed the PACSY (Protein structure And Chemical Shift spectroscopY) database, which utilizes a relational database management system (RDBMS), to manage information derived from the PDB, BMRB, and SCOP databases. We describe how information from each database is extracted and processed to make them cross-related one another to enable queries.

Materials and methods

Database design

The PACSY database was designed to store and distribute information from protein structures and NMR experiments. PACSY makes use of an RDBMS (Relational Database Management System) to implement its data submission and request features. The data are stored and maintained by the RDBMS server, and the SQL language is used for data management. An RDBMS offers advantages over a file-based database server. First, it is possible to avoid database anomalies by separating tables through database normalization (Codd 1970). In addition, data consistency can be maintained by synchronous management and parallel control and data can be standardized by organizing methods for data expression. Data integrity and recovery are additional benefits of an RDBMS database server. We developed a tool called "PACSY Maker" to create and maintain the database, and, because the SQL language is not easy to learn to operate and manage, we have developed a second tool called "PACSY Analyzer" to facilitate queries.

Figure 1 illustrates how PACSY is organized. Data from the BMRB ftp archive are acquired as a *dbmatch.csv* file. Structural information from the PDB and chemical shift information from BMRB are extracted. The PACSY Maker software then processes these data with STRIDE (Frishman and Argos 1995), combines them with SCOP data, and parses the resulting data into a set of tables and fields in the prepared RDBMS server. The data stored in the RDBMS server can be accessed by various database client application interfaces (APIs): open database connectivity (ODBC) software, Oracle's Database Express, MySQL Connector/PHP, or Microsoft's ActiveX Data Objects (ADO). The PACSY Maker program automates the building and updating of the database. It generates SQL dump files and an insertion script file.

PACSY consists of six different types of tables (Table 1). When the database is being built, PACSY Maker extracts and processes necessary information to fill these

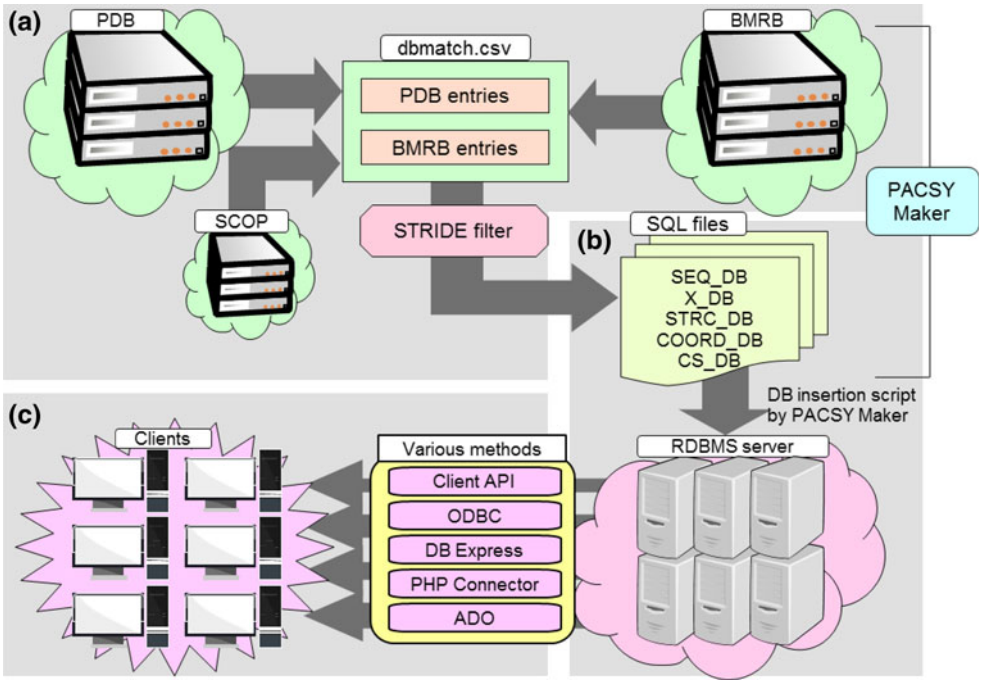

Fig. 1 Schematic representation of the PACSY database. The following steps are carried out in building and maintaining the PACSY database. **a** First, PACSY maker is used to generate SQL dump files and insertion scripts, and the dbmatch.csv file from the BMRB FTP site is analyzed to determine which entries should be incorporated into the database. **b** An RDBMS server is up before inserting the SQL files. The default insertion script is written for MySQL; however, it can be modified for other RDBMS software as long as the SQL dump files conform to the relevant SQL grammar. The settings can be optimized for the particular server environment to improve performance. **c** The database can be served by various methods supported by the RDBMS

tables from PDB, BMRB, and SCOP. STRIDE is used to calculate secondary structure type and solvent accessible surface area (SAS) (Lee and Richards 1971), and hydrophobicity scales are calculated from the SAS. The SAS values from residues are divided by those calculated from Gly-X-Gly by numerical integration to yield the relative solvent exposure. The separation of table types avoids storage of repetitive information (known as data anomalies). The "X" in front of a table type, stands for one of the 20 standard amino acids. Thus tables, *X_DB*, *X_STRC_DB*, *X_CS_DB* and *X_COORD_DB* are each actually 20 tables. Each type of table has a *KEY_ID* field. Thus, if chemical shift information about a certain residue is requested, it can be obtained by querying both the *X_CS_DB* and *X_DB* with same *KEY_ID*. Whereas other table types each consist of 20 amino-acid-specific tables, *SEQ_DB* and *SCOP_DB* are single tables. They also have a *KEY_ID* field, whose value matches that of the *X_CS_DB* for the first residue of the protein sequence.

Software design

The PACSY Maker software was developed in C++ with the Qt Developer Library (http://qt.nokia.com) for automated database generation. It builds the PACSY database by automating the flowchart shown in Fig. 2. It has the simple graphical user interface (GUI) shown in Fig. 3a, which is used to set up a working directory to store downloaded files from the PDB, BMRB, and SCOP databases along with processed files, such as SQL dump files and an insertion script file. Once a root of the working directory is set up, other directories for storage and processes are created automatically as relative directories. The user can modify those directories for more detailed setup. PACSY Maker downloads *dbmatch.csv* from the BMRB ftp archive when it is executed (Fig. 2). The file, *dbmatch.csv*, contains information on how BMRB entries are related to entries in other databases such as PDB, Swiss-Prot, and EMBL. PACSY Maker processes the file to contain only information from PDB and BMRB submitted by a common author, and checks for needed updates by comparing the results to a recently processed *dbmatch.csv* file. Next, PACSY Maker downloads the SCOP database, and parses it to add structural classification information to each PDB entry. Finally, PACSY Maker downloads PDB and BMRB files from the respective web archive that match the update list made by comparing the new and old processed *dbmatch.csv* files. Because BMRB

Table 1 Description of the six types of tables in PACSY

Table type	# of Tables	# of Fields	# of Records	Contents
SEQ_DB	1	14	7,395	Basic information for entries: sequence, pH, temp, etc
SCOP_DB	1	10	143,428	Information on the structural classification of proteins (SCOP)
X_DB	20	5	374,631	Residue-related information: e.g., chain ID, sequence ID, amino acid type
X_STRC_DB	20	9	6,098,716	Structural information for a residue: e.g., secondary structure, dihedral angles, hydrophobicity scale, SAS, # of model
X_COORD_DB	20	7	75,899,756	Coordinate information for an atom
X_CS_DB	20	5	2,035,722	Chemical shift information for an atom including assignment ambiguity

This table represents PDB and BMRB data downloaded on February 7, 2012, and data from SCOP version 1.75

has not converted fully from the old NMR-STAR v2.1 to the new NMR-STAR v3.1 file format, PACSY Maker has a parser for both file formats. PACSY Maker downloads the v3.1 file if it exists or, if not, downloads the 2.1 file. Of all the processes, this step takes the longest time, and the duration depends on the Internet bandwidth of the computer building the database. The initial run of PACSY Maker typically takes 2 h, but after the initial database creation, updates take only a few minutes. Because the PDB entry for a protein structure typically contains coordinates for multiple conformers, PACSY separates these prior to analysis by STRIDE. The model splitter module in PACSY Maker splits the downloaded PDB entry into files containing single structural models. PACSY Maker then creates output files with residues classified into six secondary structure types (H; α-helix, E; β-strand, T; turn, G; 3_{10} helix, C; coil, B; isolated β-bridge), solvent accessible surface area (SAS), and dihedral angles (PHI, PSI). PACSY Maker reads the outputs, and calculates the hydrophobicity scale of each residue from its SAS by dividing by pre-defined values of SAS of Gly-X-Gly.

The next step is to build the PACSY database. PACSY Maker generates SQL dump files and an insertion script file for RDBMS servers. First, an SQL dump file, *initdb.dmp*, is generated for initialization. It cleans existing tables and creates new tables. To (re)generate a completely new PACSY DB, the user reactivates commented-out lines in *initdb.dmp* to erase all pre-existing data. Otherwise, this file is left unedited. Second, SQL dump files containing actual data are generated: *X_DB_#.dmp*, *COORD_DB_#.dmp*, and *CS_DB_#.dmp*. The "#" characters indicate incremented indices that start from zero. For compatibility with 32-bit operating systems that handle files only smaller than 2 GB, PACSY Maker utilizes a strategy to limit file sizes. Finally, PACSY Maker generates an *insertSQL.sh* file for executing other SQL dump files. The *insertSQL.sh* file specific for MySQL has the following structure:

$$mysql - u\ USERNAME - pPASSWORD\ DBNAME < SQL\ DUMP\ FILE$$

If PostgreSQL, another popular open source database server, is used, the script file would be:

$$psql - U\ USERNAME - d\ DBNAME - f\ SQL\ DUMP\ FILE$$

Because PACSY Maker generates SQL dump files with general SQL sentences, only minimal changes are needed for field types in the *initdb.dmp*. However, field types are not always compatible between database servers. For example, MySQL requires a specific length of characters for the *TEXT* field type, whereas PostgreSQL supports variable length of *TEXT* field type. Another difference is in the nomenclature of the 8-bit floating variable: DOUBLE is used by MySQL, whereas FLOAT8 is used by for PostgreSQL. These minor changes can be easily carried out by use of any text editor.

After the *insertSQL.sh* file and *initdb.dmp* have been modified as needed, the *insertSQL.sh* can be executed for database creation. These database creation steps took 1 day for an initial run on a 2.4 GHz quad-core machine running CentOS 5.5 64 bit.

Through its interface to the PACSY RDBMS server the client software PACSY Analyzer provides an easy graphical user interface (GUI) to the PACSY database (Fig. 3b). Although the SQL language supports a powerful and standardized way to query a database, its complexity can be a barrier to non-specialists. PACSY Analyzer provides graphical user interface that allows the user to select for search tables and fields in a dialog window. Once the selections are made, PACSY Analyzer generates an SQL sentence to be executed with the PACSY database. PACSY Analyzer is written in PASCAL language (FPC version

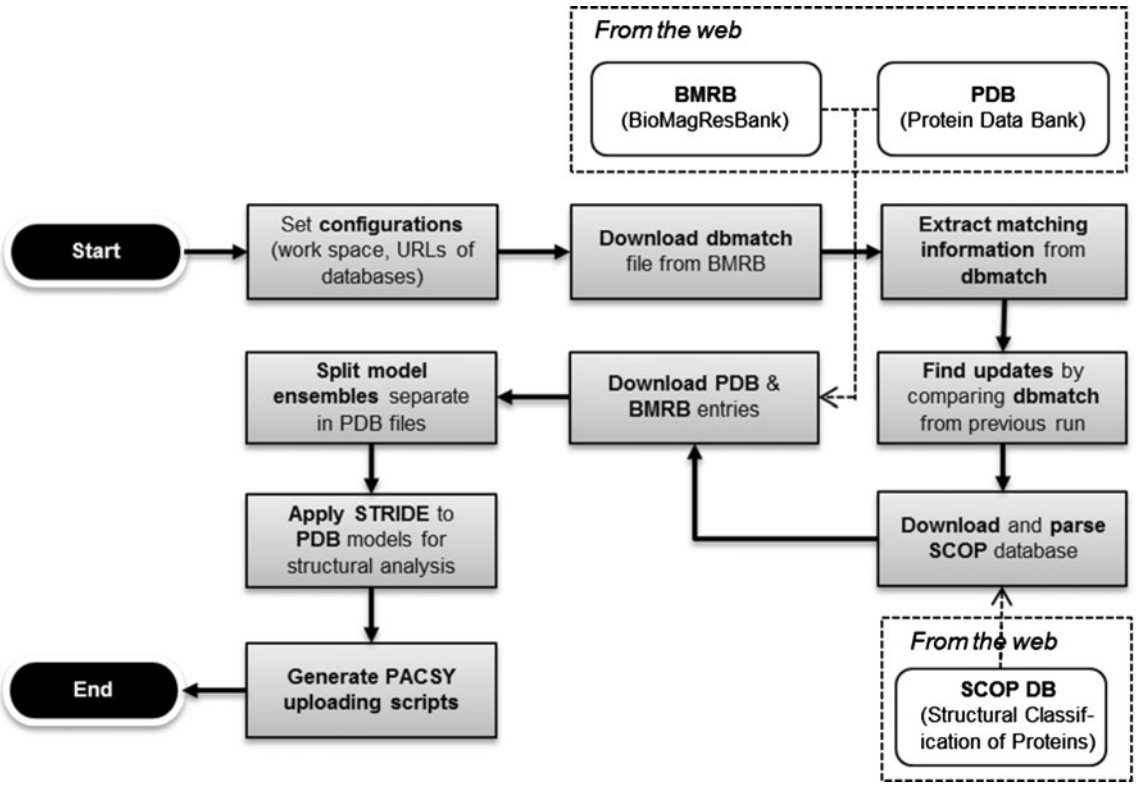

Fig. 2 PACSY maker flowchart. The PACSY maker program identifies entries to be updated and downloads them for processing. The STRIDE program provides structural information from PDB coordinates, such as dihedral angles, secondary structure, and solvent accessible surface area. The hydrophobicity scale is calculated from the solvent accessible surface area. After the downloading, processing, and updating steps, PACSY Maker generates SQL dump files and an insertion script file for the MySQL server

2.6) using Lazarus IDE (Integrated Development Environment, http://www.lazarus.freepascal.org) version 0.9.30. PACSY Analyzer supports any database server that has ODBC (Open Database Connectivity).

The dialog window has two tab controls, *Input Filters* and *Output Filters*, that are used to specify the input and output. The *Input Filters* tab sets the conditions for a search. For example, if the user wants to browse proteins whose data were collected between pH 6.0 and pH 8.0, the *pH* field in the *SEQ_DB* table is set to 6.0 and 8.0, and the "Add button" is clicked. If the user wants to search chemical shifts of only CA atoms, a filter is set in the *ATOM_NAME* field of the *X_CS_DB* table by typing CA in the text box. Filters can be set to select for any conditions supported by the PACSY database.

The *Output Filter* tab is the place where the user describes the desired output of the information to be grabbed by the *Input Filter*. If the user wants, from the above example of data collected between pH 6.0 and pH 8.0, a list of the associated PDB identifiers, the *PDB_ID* field in the *SEQ_DB* table is chosen as an output filter. From the example of a search for CA chemical shifts, if the user wants to see the mean value of chemical shifts satisfying the condition, AVG in the Statistics and C_SHIFT field of the X_CS_DB table is selected. After specifying all input and output filters, the *Make* button is clicked to create the SQL sentence that will run the user's request. The generated sentence appears in a large text box. Users can verify or edit the SQL sentence as needed to refine the search. To commit the sentence, the *Query!* button is clicked.

Depending on the SQL query, the search can take seconds or hours. Simple queries, such as browsing chemical shifts under certain conditions, are very fast (usually less than a second). The example shown Supplementary Table S2 requesting the statistics on alanine alpha-carbon chemical shifts from proteins with 80–100 residues at low pH (pH 3–5) took only 1 s. However, if the search is complex or if multiple searches are requested, more time will be required to complete the query. When multiple queries are entered, PACSY Analyzer generates a new SQL sentence after the previous one has been executed. The queried results are shown in a grid. PACSY Analyzer has a function that allows results to be exported in tab-delimited text format for use in a spreadsheet program such as Microsoft Excel or OpenOffice Spreadsheet.

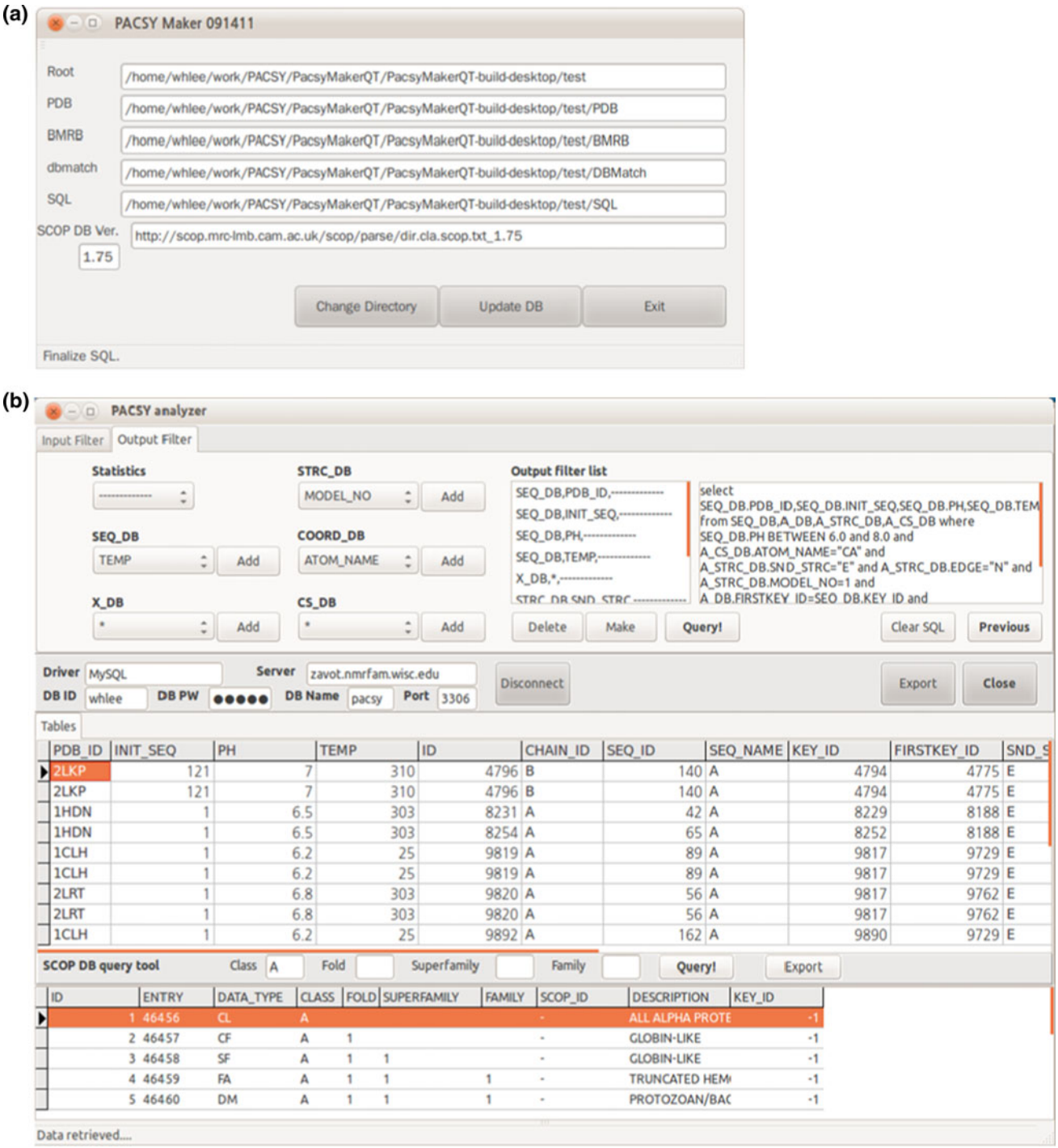

Fig. 3 Screen shots of PACSY maker and PACSY analyzer. **a** PACSY maker is a program with a simple user interface for setting up working directories. It is fully automated and does not require any user management. It takes a full day to download and process the PDB, BMRB, and SCOP databases. **b** We developed the PACSY analyzer program to provide a user interface for users not fluent in the SQL language. *Input Filter* and *Output Filter* tabs allow the used to specify the input and output PACSY queries. The output from both SQL and PACSY analyzer queries can be exported in comma-separated file format for external use

Results

Database build

The PACSY database was built and installed for testing at the National Magnetic Resonance Facility at Madison (NMRFAM). PACSY Maker ran on a 64-bit CentOS 5.5 developmental server for an entire day to build and upload SQL dump script files for the initial database. The number of downloaded PDB and BMRB files were both 3745, and a data file was downloaded for SCOP. 473 Mb were consumed by BMRB files, whereas 18 Gb were consumed by PDB files. The size of SCOP database was only 5.8 Mb.

A MySQL 5 server was installed with default parameters. The uploading process was carried out by executing the *insertSQL.sh* file after editing the user account in the *insertSQL.sh* file to change the preset *USER* and *PASSWORD* values. Execution of the *insertSQL.sh* shell script made the stored PACSY database ready for use by the MySQL server. 8460 files were generated: 648 of *X_DB_*.dmp*, 204 of *CS_DB_*.dmp,* 7590 of *COORD_DB_*.dmp*, 16 of *SEQ_DB_*.dmp*, *initdb.dmp*, *insertSQL.sh* and *update.log* files. The total size of the files was 6943 Mb (mostly SQL dump files). It took approximately 4 h to upload the PACSY data into the prepared MySQL server.

Database composition

After the files were uploaded to the server, the overall volume of PACSY storage was estimated at 5,639 Mb. Because PACSY contains only data, its size is smaller than the SQL files, which contain commands, brackets, quotes, and other SQL-related information. Of the six table types in PACSY (Table 1), the *X_COORD_DB* tables are the most space-consuming, because they contain the atom coordinates from each of the multiple conformers that represent the NMR structure of the protein as deposited in the PDB. The *X_STRC_DB* file also contains all of the structural models in the PDB entry. However, the *X_STRC_DB* file is smaller, because it contains only one record per residue, whereas the *X_COORD_DB* file contains all of the atom coordinates. The *X_STRC_DB* file has a field named *MODEL_NO*, which indicates the model number in the PDB. This makes it possible to select a particular structural model, such as the one with the lowest energy. As in SCOP and CATH, the SEQ_DB refers to chains rather than to structures; currently, 7395 chains are represented.

Nomenclature

PACSY is consistent with the IUPAC recommendations (Markley et al. 1998), which are followed by PDB and BMRB. PACSY Maker adopts the atom names from PDB and BMRB. PACSY does not use pseudo atom nomenclature; however, ambiguously assigned atoms are represented by the same chemical shift values. A field named *AMBIGUITY* in the X_CS_DB tables carries thin information; as in the BMRB, a value of 1 in the field indicates that the assignment is unambiguous, whereas a value of 2 indicates ambiguity.

PACSY statistics

Statistics were collected from PACSY to confirm both the availability and feasibility of database queries. Because the PACSY database employs a client–server concept, it supports many different options, including remote operation (Fig. 1). Because PACSY Analyzer utilizes an ODBC connection to the database server, in our case MySQL 5.0, we first installed and set up ODBC Connector. Next, we used PACSY Analyzer to determine the structural classification of PACSY entries as defined by the SCOP database (Table 2). SCOP does not cover all PDB entries, because full classification is not automated. Csaba's study in 2009 revealed that the SCOP database version 1.73 covered 35.5 % of all PDB entries whereas CATH database version 3.1.0 covered 32.0 % (Csaba et al. 2009). Furthermore, Jefferson and co-workers found that for single domain classifications of the type commonly found in NMR structures, coverage of CATH by SCOP was greater than that of SCOP by CATH (Jefferson et al. 2008). We found that the SCOP 1.73 database provided 43 % coverage. Because PACSY contains structural classification information, it is possible to investigate proteins by fold class. Apart from unclassified entries, the largest class of PDB and BMRB entries were for all-alpha proteins (745 entries, Table 2). Other major classes are well represented, except for multi-domain proteins (no entries, Table 2).

We also determined the mean and standard deviation values of the chemical shifts of the backbone atoms ($^{13}C^{\alpha}$, $^{13}C'$, ^{15}N, ^{1}H, $^{1}H^{\alpha}$) of the 20 standard amino acids as a function of 6 secondary structure types. The values were calculated by a short Python script. Strong relationships between local structure and chemical shifts are known to exist (Han et al. 2011; Iwadate et al. 1999; Kohlhoff et al. 2009; Meiler 2003; Moon and Case 2007; Vila et al. 2009). Thus, we expected to see distinct chemical shift differences between secondary structures, particularly three major secondary structure types, α-helix, β-strand, and random coil residues. Figure 4a and Supplementary Table S1a shows results for the alpha carbon ($^{13}C^{\alpha}$) chemical shifts. To visualize the distinctions between amino acid and structure types, we calculated differences for each of the 6 structure types from the average over all 6 (Fig. 4b). Statistics for the chemical shifts of the four other backbone

Table 2 PACSY as classified by the SCOP database

Class	Number of entries
A—All alpha proteins	745
B—All beta proteins	555
C—Alpha and beta proteins (A/B)	580
D—Alpha and beta proteins (A + B)	443
E—Multi-domain proteins (alpha and beta)	0
F—Membrane and cell surface proteins and peptides	14
G—Small proteins	467
H—Coiled coil proteins	18
I—Low resolution protein structures	1
J—Peptides	166
K—Designed proteins	99
Unassigned	4,307
Total	7,395

This table represents PDB and BMRB data downloaded on February 7, 2012, and data from SCOP version 1.75

atom types are also in Supplementary Table S1. The results showed that mean chemical shifts differ by amino acid type, atom type, and secondary structure class, α-helix (H), β-strand (E), or coil (C). For example, the mean ^{15}N chemical shifts of Ala in α-helical (121.72 ppm) and coil (124.48 ppm) environments differ by 2.76 ppm, whereas those for β-strand (124.85 ppm) and coil (124.48 ppm) differ by only 0.37 ppm. By contrast, the mean ^{15}N chemical shifts of Thr in α-helical (114.86 ppm) and coil (114.99 ppm) environments differ by only 0.13 ppm, whereas those for β-strand (117.70 ppm) and coil (114.99 ppm) differ by 2.71 ppm. This kind of analysis can be refined by any of the conditions available in the PACSY database, e.g., pH, temperature, hydrophobicity scale or solvent accessible surface area (SAS).

Practical example

We provide an example of how PACSY can be used in practice (Supplementary Table S2). We assume that a novel protein of interest contains 90-residues and has been shown to be all α-helical and stable at low pH. As an aid to assignment, we are interested in knowing the range of $^{13}C^{\alpha}$ chemical shifts for an alanine residue under these conditions and how the chemical shift may depend on backbone torsion angles (PHI, and PSI), solvent accessible surface (SAS), and hydrophobicity scale (HDO_PBT).

The conditions to be searched were inserted into the Input Filter tab of PACSY Analyzer (Supplementary Table S2a). To limit the size of proteins to those near 90 residues, we set the residue number (SEQ_COUNT) to "80" (minimum) and "100" (maximum); to limit the output to alanine residues, we set CLASS to "A"; to include a range of low pH values we set PH to "3"(minimum) and "5" (maximum); because we were not interested in comparing multiple conformers representing solution structures, we set MODEL_NO to "1"; since we were interested in all helical proteins, we set SND_STRC to "H" and EDGE to "N" (no mixed secondary structure); to limit the atom queried to Cα, we set ATOM_NAME to "CA".

We specified the desired output data in the Output Filter tab of PACSY Analyzer (Supplementary Table S2b). Items requested from the sequence database (SEQ_DB) were: the PDB structure identifier (PDB_ID); the chain designator (for proteins containing more than one peptide) (CHAIN_ID); the BMRB, accession number (BMRB_ID), the total number of residues in the chain (SEQ_COUNT), the pH (PH) and temperature (TEMP) at which the NMR data were acquired. Requested from the chain database (X_DB) was the residue number (SEQ_ID) for the particular chain (CHAIN_ID). Items requested from the structure database (STRC_DB) were the ϕ (PHI) and ψ (PSI) torsion angles, the hydrophobicity scale (HDO_PBT), and the solvent accessible surface (SAS). The only request from the chemical shift database was the ^{13}C chemical shift (C_SHIFT).

PACSY Analyzer automatically generated the SQL sentence to be submitted to the PACSY database (Supplementary Table S2c). The advanced search performed by the PACSY database took less than 10 s. The *Export* feature of PACSY Analyzer was used to save the result in comma-separated value (.csv) format for input into a spreadsheet or text editing program (Supplementary Table S2d).

The search of the PDB, BMRB and SCOP databases identified 50 alanine residues in six proteins (PDB ID: 1HUE, 1AAB, 1QPU, 1XSX, 2JN6, 2JS1). Whereas the $^{13}C^{\alpha}$ chemical shifts in the full BMRB have a mean value of 53.2 ppm and a standard deviation of 2.4 ppm range, this restricted search yielded a mean value of 53.8 with a standard deviation of 1.30. The results could be filtered further, for example on the basis of ϕ,ψ angles.

We wrote a short Python script to show another practical application of PACSY (available from the PACSY website). We used PACSY to determine the correlation between chemical shifts and hydrophobicity scale values. Figure 5 shows how chemical shift and hydrophobicity are related in for alanine residues. In addition, the trend is also dependent on secondary structure type. If the secondary structure was α-helix, the chemical shift tended to increase when the residue was more exposed to the solvent. On the other hand, if the secondary structure was β-strand, the chemical shift tended to decrease when the residue was more exposed. The exposure rate predicted from the

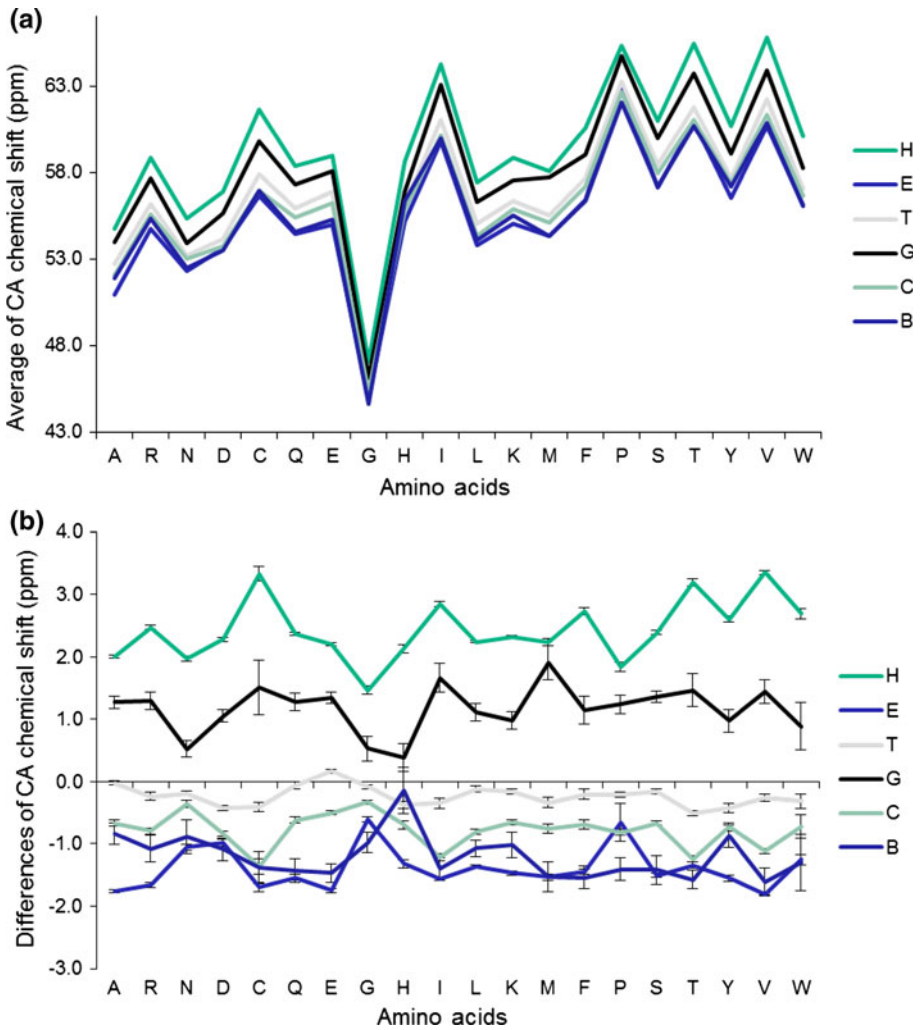

Fig. 4 Mean $^{13}C^{\alpha}$ chemical shifts for different types of amino acids according to different classes of secondary structure retrieved from the PACSY database. We wrote a short Python script (available from the PACSY website) to collect chemical shift statistics on 5 major backbone atoms in the PACSY database only one of which is plotted here. The abbreviations for secondary structure classes are: α-helix (H), β-strand (E), turn (T), 3_{10} helix (G), coil (C), and isolated β-bridge (B)

chemical shift could be used as an added target function to structure calculations.

Conclusions

PACSY introduces a way of both storing and categorizing structural and chemical shift data. It supports easy data queries based on information from the PDB, BMRB, and SCOP databases. To create this environment, we first defined the database structure and table descriptions; then we created the PACSY Maker program that automatically downloads, parses, processes, and stores data from PDB, BMRB, and SCOP. PACSY Analyzer was designed to make the PACSY database accessible to users without experience in creating SQL queries. PACSY Analyzer has a graphical user interface and an automatic SQL generating function. As an initial test of the PACSY database, we carried out a query that returned the dependence of protein backbone chemical shifts on amino acid residue type and classification of their secondary structure (Supplementary Table S1). The script used in that query is available from the PACSY website (http://pacsy.nmrfam.wisc.edu). We also show an example of how PACSY Analyzer can be used to generate a complex SQL query.

PACSY will enable research focused on the relationship between local structure and chemical shifts. Studies can employ as variables, temperature, pH, SCOP class, or sequence length. PACSY is easily extensible because it makes use of an RDBMS server. Users can make use of powerful SQL queries to edit the PACSY database for specific purposes. If a new feature needs to be added, the JOIN or ALTER commands can be used to modify table structures or to add another field. If an added feature is quite distinct from pre-existing tables, the table can be included in PACSY by specifying a KEY_ID field that refers to the PACSY database. We envision that PACSY will be found useful as a tool for assisting NMR peak assignments as illustrated by the practical example. Researchers interested in protein structure prediction from chemical shifts can filter the PACSY database to test hypotheses. These can involve coordinates from the

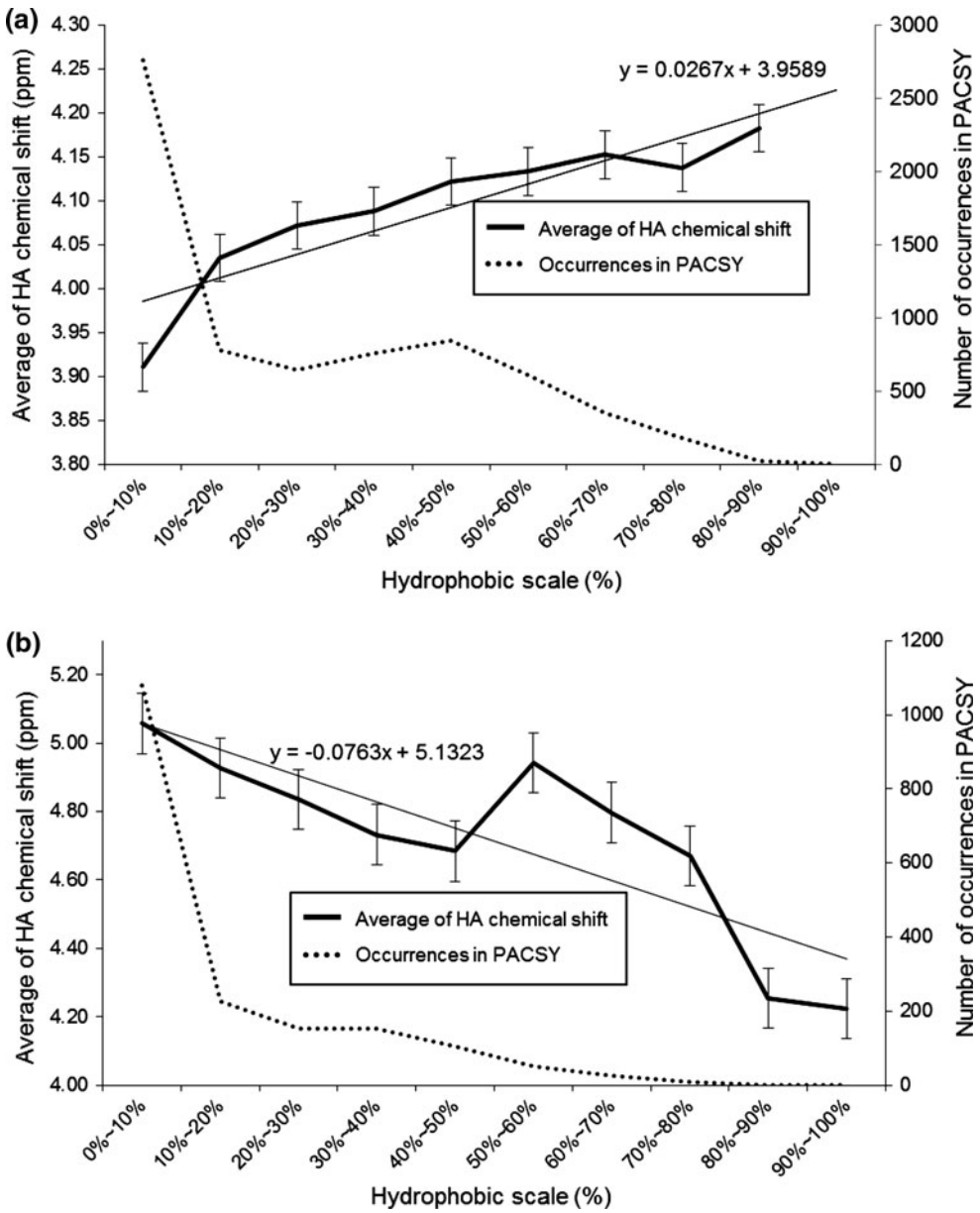

Fig. 5 Mean alanine $^{1}H^{\alpha}$ chemical shifts as a function of the hydrophobicity value. The short Python script used to acquire the text data from PACSY and to draw this plot is available from the PACSY website. **a** Double *y*-axis plot showing the mean alanine $^{1}H^{\alpha}$ chemical shifts for residues in α-helix and the number of occurrences as a function of hydrophobicity. **b** Double *y*-axis plot drawn showing mean alanine $^{1}H^{\alpha}$ chemical shifts for residues in β-strand the and number of occurrences as a function of hydrophobicity

X_COORD_DB, structure types from the X_STRC_DB, sequence information from SEQ_DB, and chemical shift information from X_CS_DB. PACSY also can be used as a NOESY simulator for known structures. Coordinates of hydrogens closer than 5 Å can be searched from the PACSY database, and the matching chemical shifts from the X_CS_DB table can be assembled to simulate NOESY.

The PACSY website (http://pacsy.nmrfam.wisc.edu) accepts SQL command line requests from users. Users unfamiliar with SQL can use PACSY Analyzer to generate SQL commands. For those who wish to build their own PACSY database, executable files for PACSY Maker and PACSY Analyzer are available from the website.

Acknowledgments This work was supported by NIH grants from the National Center for Research Resources (5P41RR002301-27 and RR02301-26S1) and the National Institute for General Medical Sciences (8 P41 GM103399-27) that support the National Magnetic Resonance Facility at Madison (NMRFAM). W.Y. and I.C. were supported by the Creative Research Initiative (No. 2011-0000041) to Center for Proteome Biophysics from National Research Foundation/Ministry of Education, Science and Technology, Korea. We thank Dimitri Maziuk for technical support of the NMRFAM computer network. All the work performed in this paper is based on the versions of PDB and BMRB available on February 7, 2012 and on the 1.75 version of the SCOP database.

References

Bahrami A, Tonelli M, Sahu SC, Singarapu KK, Eghbalnia HR, Markley JL (2012) Robust, integrated computational control of NMR experiments to Achieve Optimal Assignment by ADAPT-NMR. PLoS ONE 7:e33173

Bardiaux B, Malliavin T, Nilges M (2012) ARIA for solution and solid–state NMR. Methods Mol Biol 831:453–483

Berjanskii MV, Neal S, Wishart DS (2006) PREDITOR: a web server for predicting protein torsion angle restraints. Nucleic Acids Res 34:W63–W69

Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 35:D301–D303

Bernstein FC, Koetzel TF, Williams GJB, Meyer EF Jr, Brice M, Rogers JR, Kennard O, Shimanouchi T, Tasumi M (1977) The Protein Data Bank: a computer-based archival file for macromolecular studies. J Mol Biol 112:535–542

Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL (1998) Crystallography and NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 54:905–921

Codd EF (1970) A relational model of data for large shared data banks. Commun ACM 13:377–387

Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 13:289–302

Csaba G, Birzele F, Zimmer R (2009) Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. BMC Struct Biol 9:23

Eghbalnia HR, Bahrami A, Tonelli M, Hallenga K, Markley JL (2005a) High-resolution iterative frequency identification for NMR as a general strategy for multidimensional data collection. J Am Chem Soc 127:12528–12536

Eghbalnia HR, Wang L, Bahrami A, Assadi A, Markley JL (2005b) Protein energetic conformational analysis from NMR chemical shifts (PECAN) and its use in determining secondary structural elements. J Biomol NMR 32:71–81

Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. Proteins 23:566–579

Gledhill JM Jr, Wand AJ (2012) Al NMR: a novel NMR data processing program optimized for sparse sampling. J Biomol NMR 52:79–89

Güntert P (2004) Automated NMR structure calculation with CYANA. Methods Mol Biol 278:353–378

Hall SR, Spadaccini N (1994) The STAR File: detailed Specifications. J Chem Inf Comput Sci 34:505–508

Han B, Liu Y, Ginzinger SW, Wishart DS (2011) SHIFTX2: significantly improved protein chemical shift prediction. J Biomol NMR 50:43–57

Hiller S, Fiorito F, Wuthrich K, Wider G (2005) Automated projection spectroscopy (APSY). Proc Natl Acad Sci USA 102: 10876–10881

Hyberts SG, Takeuchi K, Wagner G (2010) Poisson-gap sampling and forward maximum entropy reconstruction for enhancing the resolution and sensitivity of protein NMR data. J Am Chem Soc 132:2145–2147

Iwadate M, Asakura T, Williamson MP (1999) C alpha and C beta carbon-13 chemical shifts in proteins from an empirical database. J Biomol NMR 13:199–211

Jefferson ER, Walsh TP, Barton GJ (2008) A comparison of SCOP and CATH with respect to domain–domain interactions. Proteins 70:54–62

Kim S, Szyperski T (2003) GFT NMR, a new approach to rapidly obtain precise high-dimensional NMR spectral information. J Am Chem Soc 125:1385–1393

Kohlhoff KJ, Robustelli P, Cavalli A, Salvatella X, Vendruscolo M (2009) Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. J Am Chem Soc 131:13894–13895

Lee BM, Richards FM (1971) The interpretation of protein structures: estimation of static accessibility. J Mol Biol 55:379–400

Markley JL, Bax A, Arata Y, Hilbers CW, Kaptein R, Sykes BD, Wright PE, Wüthrich K (1998) Recommendations for the presentation of NMR structures of proteins and nucleic acids (IUPAC recommendations 1998). Pure Appl Chem 70:117–142

Markley JL, Ulrich EL, Berman HM, Henrick K, Nakamura H, Akutsu H (2008) BioMagResBank (BMRB) as a partner in the Worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions. J Biomol NMR 40:153–155

Meiler J (2003) PROSHIFT: protein chemical shift prediction using artificial neural networks. J Biomol NMR 26:25–37

Moon S, Case DA (2007) A new model for chemical shifts of amide hydrogens in proteins. J Biomol NMR 38:139–150

Murzin AG, Brenner SE, Hubbard T, Chothia CH (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH—a hierarchic classification of protein domain structures. Structure 5:1093–1108

Schulte-Herbruggen T, Briand J, Meissner A, Sorensen OW (1999) Spin-state-selective TPPI: a new method for suppression of heteronuclear coupling constants in multidimensional NMR experiments. J Magn Reson 139:443–446

Schwieters CD, Kuszewski JJ, Tjandra N, Marius CG (2003) The Xplor-NIH NMR molecular structure determination package. J Magn Reson 160:65–73

Sgourakis NG, Lange OF, DiMaio F, Andre I, Fitzkee NC, Rossi P, Montelione GT, Bax A, Baker D (2011) Determination of the structures of symmetric protein oligomers from NMR chemical shifts and residual dipolar couplings. J Am Chem Soc 133:6288–6298

Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci USA 105:4685–4690

Shen Y, Delaglio F, Cornilescu G, Bax A (2009a) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44:213–223

Shen Y, Vernon R, Baker D, Bax A (2009b) De novo protein structure generation from incomplete chemical shift assignments. J Biomol NMR 43:63–78

Stanek J, Kozminski W (2010) Iterative algorithm of discrete Fourier transform for processing randomly sampled NMR data sets. J Biomol NMR 47:65–77

Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent Wenger R, Yao H, Markley JL (2008) BioMagResBank. Nucleic Acids Res 36:D402–D408

Vila JA, Arnautova YA, Martin OA, Scheraga HA (2009) Quantum-mechanics-derived 13Calpha chemical shift server (CheShift) for protein structure validation. Proc Natl Acad Sci USA 106: 16972–16977

Wishart DS, Sykes BD (1994) The ^{13}C chemical shift index: a simple method for the identification of protein secondary structure using ^{13}C chemical shifts. J Biomol NMR 4:171–180

Wishart DS, Sykes BD, Richards FM (1992) The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. Biochemistry 31:1647–1651