Journal of Biomolecular NMR, **23:** 85–102, 2002.
KLUWER/ESCOM

PESCADOR: The PEptides in Solution ConformAtion Database: Online Resource

Anne Pajon[a,*], Wim F. Vranken[a,*,**], Maria Angeles Jimenez[b], Manuel Rico[b] & Shoshana J. Wodak[a,***]
[a]*Service de Conformation des Macromolécules Biologiques, Université Libre de Bruxelles, Brussels, Belgium;* [b]*Instituto de Estructura de la Materia, Consejo Superior de Investigaciones Científícas, Madrid, Spain*

Received 3 January 2002; Accepted 29 April 2002

Key words: CD, database, NMR, peptide conformation

Abstract

In recent years a large body of data has been obtained from Nuclear Magnetic Resonance and Circular Dichroism experiments on the influence of the amino acid sequence and various other parameters on the conformational state of peptides in solution. Interpreting the experimental data in terms of the conformational populations of the peptides remains a key problem, for which current solutions leave appreciable room for improvement. Considering that making this body of data available for surveys and analysis should be instrumental in tackling the problem, we undertook the development of Pescador: The 'PEptides in Solution ConformAtion Database: Online Resource'. Pescador contains data from NMR and CD spectroscopy on peptides in solution as well as information on the structural parameters derived from these data. It also features specialized Web-based tools for data deposition, and means for readily accessing the stored information for analysis purposes. To illustrate the use of the database in deriving information for the conformational analysis of peptides, we show how the alpha proton δ-values stored in Pescador and measured by NMR for different peptides in different laboratories can be used to derive a new set of 'random coil' chemical shift values. Firstly, we show these values to be very similar to those obtained experimentally for model peptides in water, and their variation with increasing Tri-Fluoro-Ethanol (TFE) concentration is similar to that reported for model peptides. We show, furthermore, that the chemical shift data in Pescador can be used to derive correction factors that take into account effects of neighboring residues. These correction factors compare favorably with those recently derived from a series of model GGXGG peptides (Schwarzinger et al., 2001). These encouraging results suggest that, as the quantity of NMR data on peptide deposited in Pescador increases, surveys of these data should be a valuable means of deriving key parameters for the analysis of peptide conformation.

Introduction

Over the past decade a large number of studies have been devoted to the analysis of peptide conformational preferences and interactions in solution. Several of these used peptides with designed amino-acid sequences in order to elucidate the factors governing secondary structure formation. A large body of work was devoted to α-helix formation and conformational equilibrium (Chakrabartty et al., 1991; Padmanabhan and Baldwin, 1994; Baldwin, 1995; Muñoz et al., 1995). With increasing interest in aggregation phenomena, believed to involve transitions from helical to extended conformation (Lopez-Hernandez and Serrano, 1995, 1996; Taddei et al., 2000; Andersen and Tong, 1997), the focus has recently shifted towards peptides forming β-hairpins and small β-sheets (Lacroix et al., 1999; Odaert et al., 1999; Santiveri et al., 2000, 2001; Gellman, 1998; Zerella et al.,

*Both authors contributed equally to this work.
**Current address: European Bioinformatics Institute, Cambridge, U.K.
***To whom correspondence should be addressed. E-mail: shosh@ucmb.ulb.ac.be

2000; Griffiths-Jones et al., 1999; Griffiths-Jones and Searle, 2000). Other studies on peptides concentrated on the local sequence dependence of NMR parameters such as chemical shifts (Bundi and Wüthrich, 1979; Merutka et al., 1995; Wishart et al., 1995; Schwarzinger et al., 2001) and coupling constants (Millhauser et al., 1996; Griffiths-Jones et al., 1998), which provide information on the secondary structure preferences, their relative populations, and the influence of local sequence.

This large body of NMR and CD data from which conformational parameters of peptides have been deduced is, however, not readily available because it remains in possession of the individual research laboratories in which the various studies were performed. Only a small portion of data appears in print in scientific literature.

Here we describe an effort to assemble and organize this information in a dedicated relational database, Pescador: The 'PEptides in Solution ConformAtion Database: Online Resource'. This database contains the data obtained from NMR and CD spectroscopy experiments on peptides in solution as well as information on the structural parameters derived from these data. It also features specialized Web-based tools for data deposition, means for readily accessing the stored information and for performing various relevant analyses. The database can be accessed on the World-Wide Web at the address: http://www.ucmb.ulb.ac.be/Pescador/

Furthermore, we show how the database can be used to derive useful information for the conformational analysis of peptides. To that end we present the analysis of the alpha proton δ-values measured by NMR, a type of readily and abundantly collected data, which, unlike other types of data, are available for over nearly all the peptides (about 145 in total) stored in Pescador. Since a wide range of analyses on peptides and their conformation concentrates on the proton δ-values, those values are an ideal starting point for evaluating the application of Pescador. We show how the chemical shift data and experimental information can be used to derive a new set of reference values for alpha proton chemical shifts of amino acids. This set is based on peptides with heterogeneous sequences that have no known conformational preference. In addition, sequence neighbor effects are examined within this set in order to derive correction factors which allow the calculation of 'preferred random coil' values based on sequence rather than 'random coil' values alone. Another effect analyzed is that of TFE on the amide proton δ-values. The values obtained and trends observed here are compared to those derived from experimental measurements on individual peptides, and are shown to be in the expected range. This is, to the best of our knowledge, the first time that reference values for chemical shifts are obtained from surveys of pooled experimental data from different peptides. These encouraging results suggest that, as the quantity of curated experimental data on peptide stored in Pescador increases, surveys of this kind should be a valuable means of deriving key parameters for the analysis of conformational preferences of peptides in solution.

Materials and methods

Database schema, and organizations

In the relational database Pescador, each database entry represents one experiment, having a unique identifier (see Figure 1). This experiment is characterized by a set of experimental conditions (Table 'Experimental_conditions' with pH, temperature, ... Table 'Solvents' with solvent type and solvent percent, and Table 'AddComponents' with the additional components.). There are four other sections associated with each experiment. The first section is the global information part, which contains a description of the source of the data (relevant literature references in Tables 'Refs' and 'Authors' and laboratory information in Table 'Labs'). The second section contains all the information necessary for the full characterization of the chemical structure of the peptide. This includes the specification of non-natural amino acids, residues with cis peptide bonds (Table 'CisResidues') and additional covalent bonds (Table 'SupplementaryCovBonds'). The third section houses the primary experimental CD and/or NMR data, measured at the specific experimental parameters and conditions specified in the first section. Depending on the type of data, values are stored for individual residues or individual atoms. The last section of the database contains information on the peptide conformation. This includes global populations (Table 'Global_population'), secondary structures (Table 'QualStructAnalValues') and hydrogen bonds (Table 'HBondsValues') as determined from the primary data by the experimentators, as well as average angles (Table 'AverageAnglesValues') and calculated structures (Table 'CalculatedStructures'). The Pescador database contains a total of 40 Tables, and is implemented using the SYBASE RDBMS.

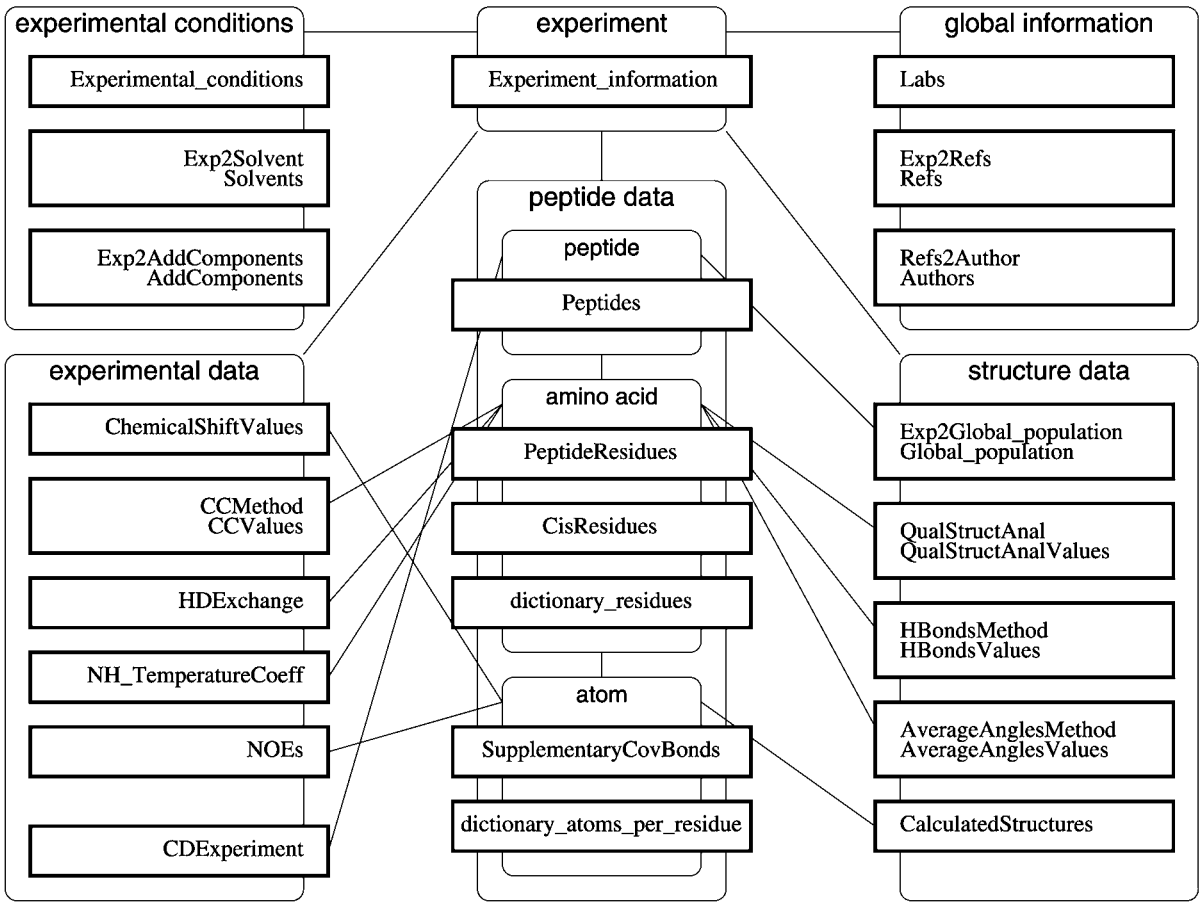

Figure 1. Pescador Database schema. Each bold rectangle represents a Table containing certain properties or attributes. Relationships between Tables are indicated by connecting lines. Rounded rectangles represent sections: virtual groups of Tables that are associated by their data type. The five sections are centered around the experiment Table. In the section experimental data, the Table 'ChemicalShiftValues' contains the δ-values, 'CCMethod' and 'CCValues' contains coupling constant information, 'HDExchange' contains values from H/D exchange experiment, the Table 'NH_TemperatureCoeff' contains the NH chemical shift temperature coefficients, 'NOEs' contains NOE data and the Table 'CDExperiment' contains experimental CD data. The explanation of all the other ones can be found in the paragraph 'Database schema, and organizations'.

Database integrity

To ensure the correctness of data in the database, a large number of constraints is introduced. These are of three types. One is the obligatory primary key, which is a unique identifier for a Table (Figure 2, grey rectangles). The second is the foreign key which represents a reference to the matching columns of a row in a target Table (Figure 2, dotted lines). This ensures that the value of the foreign key and the value of the primary key in a row of the target Table are identical. The third type is a check, specific to each column, for example, to ensure that the value of an entered parameter, such as pH, is within an authorized range (in this case between 1–14). To illustrate the importance of the constraints, Figure 2 shows an example based on the 'ChemicalShiftValues' Table. The entries in this Table are δ-values of each nucleus in each experiment. The primary key is the couple of values, 'ExpID' and 'Nucleus_shift_assign_ID', which is different for each row in the Table. There are three foreign keys, which define links to other Tables, which are the 'Experimental_information', the 'PeptideResidues' and 'dictionary_atoms_per_residue' Tables. The foreign key consisting of the couple of columns ('ExpID', 'PeptID') provides such link to the 'Experiment_information' target Table. This ensures that δ-data are entered only after crucial information about the experiment is already present in the database. Similarly, to enter the δ-value of a nu-

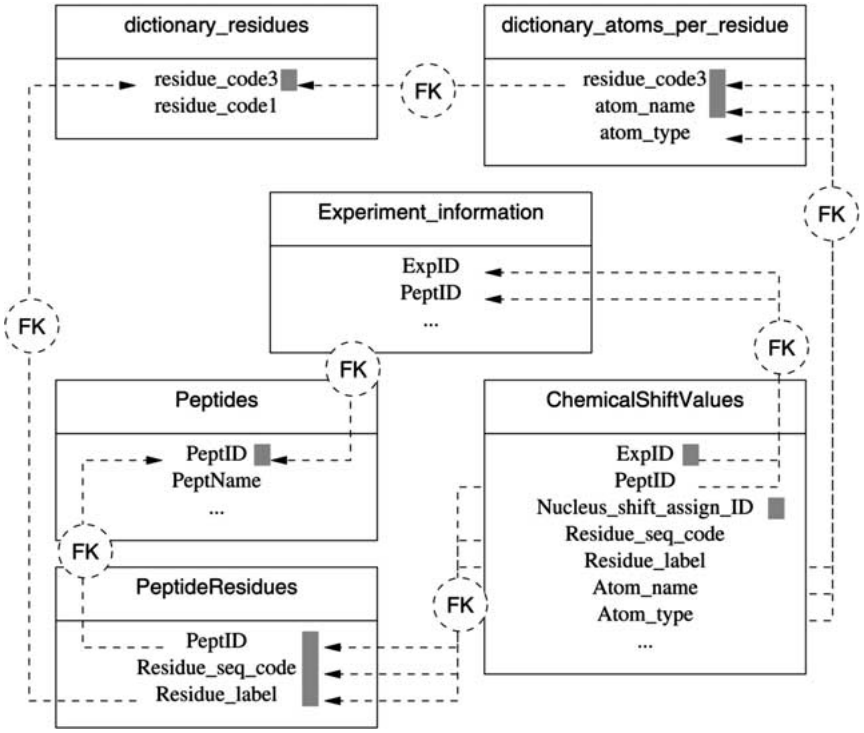

Figure 2. Schema of the constraints on the chemical shift Table. Each rectangle is a Table, and dotted lines between Table columns designate foreign keys. The grey rectangles represent the unique key for each Table.

cleus belonging to a given residue, the residue name and number must already be present in the 'PeptideResidues' Table and the atom associated with the residue must be in the 'dictionary_atoms_per_residue' Table, which contains the atom list of each residue.

Web based deposition tools

A key aspect for any database is the ease with which data can be deposited. To facilitate data deposition in Pescador, we therefore developed a web-based deposition tool (see Figure 3). This tool facilitates and speeds up data entry, through forms and selection lists (see Figure 3). The different sections of the deposition form are organized as HTML pages which can be easily accessed through a clickable 'Table of contents' (left-hand side of Figure 3). Entries in these pages are gradually flagged as each step of the deposition process is completed.

Upon submission, a specific ID number is assigned to the deposition. A simple search engine allows the user to extract ID numbers from Pescador based on a set of search parameters. To avoid loss of data, raw data entered are directly saved in a temporary file, and certain simple checks are then performed before the data are stored in the database. These checks are carried out on the computer of the depositor using Javascript programs (top grey square of Figure 4). They consist mainly in verifying that the mandatory fields have been filled. For example, when the number of solvent components entered is three, like in Figure 3, the program checks that the depositor has filled in the three lines in the solvent Table below, and the two columns containing information on the type of solvent used and their respective proportions in the solution. Only then will the deposition step for the experimental conditions be completed and the corresponding flag will turn green.

For certain parameters (e.g., chemical shift data) the depositor can enter the data from a file in one of several common formats, such as InsightII.ppm, noe2pdb.par, XEasy.prot, NMR-STAR, Pronto report, PDB-Pipp, Sparky.proj and PENCE. Automatic 'Chemical shift deviation' graphs can be obtained based on the most common reference values (Bundi and Wüthrich, 1979; Merutka et al., 1995; Wishart et al., 1995). Such graphs can be computed on the fly during deposition, when the chemical shift data are

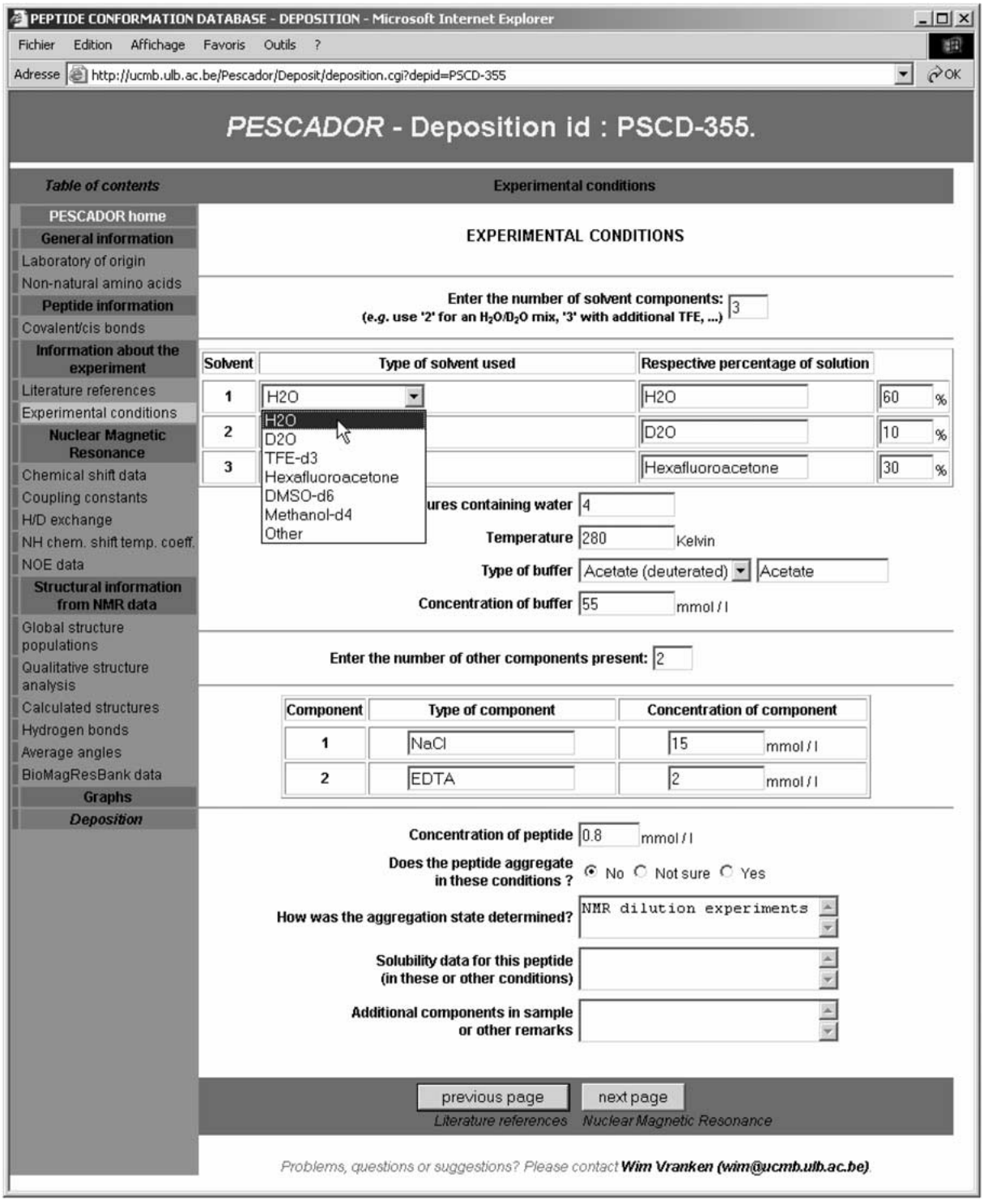

Figure 3. Overview of a typical Pescador deposition form.

entered manually, or at any time after these data have been entered from file and properly checked.

When all the flags turn green, the 'Deposition' page (bottom of 'Table of contents') appears and

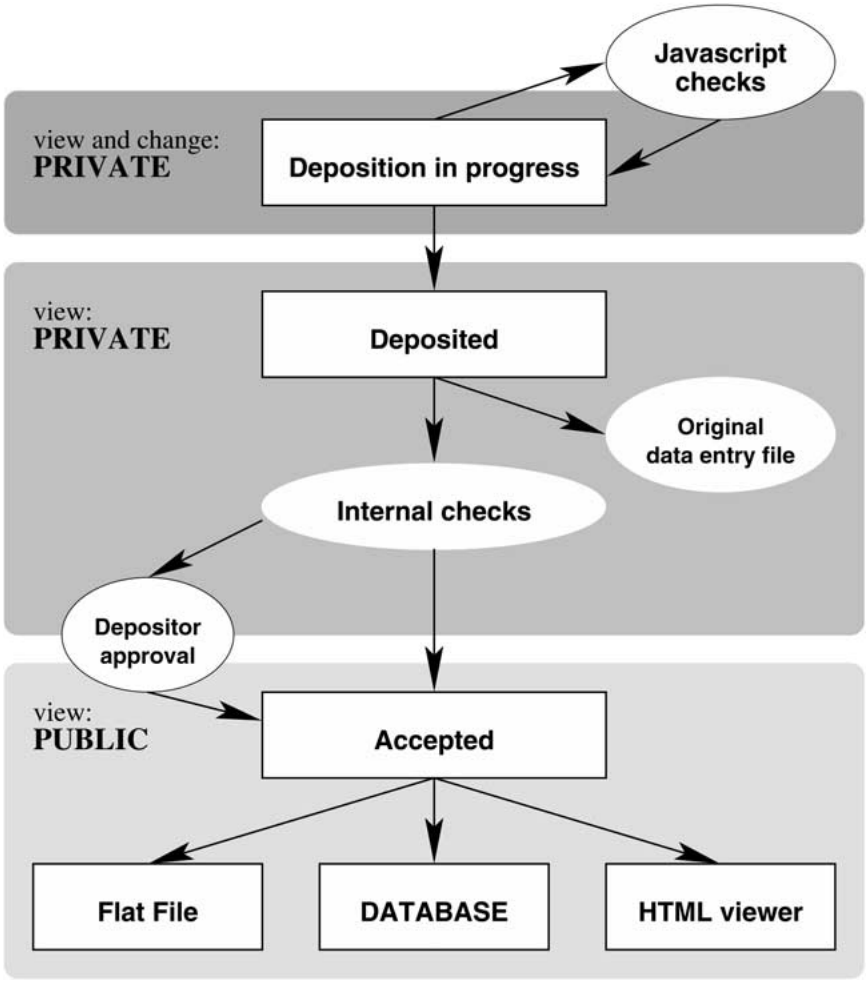

Figure 4. Flow chart for data processing in Pescador. Each grey rounded rectangle represents a different accessibility of the data (private or public). The data states are surrounded by bold rectangle. Circles surrounded by two arrows modify data states. Arrows show the flow of the data.

the data can be committed to Pescador (middle grey square of Figure 4). During the entire deposition process, access and deposition of the data are password protected (Figure 4): Data viewing is private.

Data processing

A procedure has been developed for processing the data after submission (middle and bottom grey squares of Figure 4). This procedure operates on a separate copy of the data. It starts by standardizing the data and then scans them in order to detect errors. If a chemical shift file was deposited, the standardization consists in changing the atom names to the internally used IUPAC naming (Markley et al., 1998) whenever necessary and adding stereocodes to obtain chemical shift data in a form ready to be entered into the various Pescador Tables. Then, values of many data items (such as Journal and laboratory names, names of peptides, solvents and amino-acid sequences) are compared against a dictionary that contains the 'standard' values, defined as those compiled previously from reference sources, or already stored once or more in the database. Values with no match in the dictionary are flagged for further examination by the database curator and in some cases the depositor. A submission is accepted only after either one, or in some cases both, have given their approval (bottom grey square of Figure 4).

At the end of the processing procedure the data is converted into a text format similar to NMR-STAR used by BMRB (Seavey et al., 1991) and then entered into the Pescador Tables.

Options are provided to the depositor to download the data as an NMR-STAR file and to make a concomitant deposition to BMRB. In the latter case the depositor is prompted additional information required only by BMRB.

Data acquisition

Acquiring a minimum body of data is necessary before a database such as this one can be properly validated and considered of use to the scientific community. To achieve this goal, we entered into the database the set of NMR data on peptides previously collected for developing the Agadir program (Muñoz and Serrano, 1994; Muñoz and Serrano, 1995a,b; Lacroix et al., 1998) at the EMBL-Heidelberg. In addition, a large set of peptide NMR data available at the Instituto de Estructura de la Materia (IEM) was processed and deposited in Pescador. Those were complemented by data available from the literature. Since Pescador focuses on peptides and their conformations, a limit of 30 residues was set for the length of the peptides accepted for deposition in Pescador.

Data retrieval and searching

Data browsing, full database downloads and keyword queries and searches are available from the main web page of Pescador. The link 'View accepted data' offers the facilities to look at the accepted data in the PESCADOR table format or to download the full file in a STAR format. Graphs showing the deviation from random coil values (from several sources) are offered for the alpha protons and the alpha carbons, and postscript versions of these graphs can be downloaded. A simple search facility is available under the link 'Search PESCADOR entries'. Different fields like laboratory information, bibliographic references, solvent, temperature, pH and peptide sequence can be queried.

Query and analysis capabilities

In order to analyze deposited data we developed SQL scripts that automatically produce surveys of various parameters. This includes distributions of the stored values and various parameters characterizing these distributions such as averages, median, minimum and maximum values and standard deviations. In addition, database views were generated in which particular ranges of values or types of data are selected. For example, to examine the properties of peptides without a stable conformation, we defined a restricted subset of the data, from which peptides having a global secondary structure population over 20% and amino acids reportedly part of a secondary structure element were excluded. Excluding peptides whose conformation was studied in solvents other than aqueous solvent (H_2O and/or D_2O), or residues positioned at the peptide extremities is also enabled by using the right combination of SQL queries.

The output of the queries was generated in a format readable by the program R, a language and environment for statistical computing and graphics (Ihaka and Gentleman, 1996). We implemented several R scripts to automatically generate analysis results and graphical representations of the data.

Results

Overview of data currently deposited

An overview of the data stored in Pescador is presented in Table 1 and Figure 5. The database presently contains 233 deposited and processed experiments for a total of 145 peptides, whose average length is 16.97 residues and the distribution is depicted in Figure 5a. These data originate from 15 different laboratories, with 136 depositions originating from the IEM, and several depositions based on data available from literature. The distribution according to solvent type (Figure 5c) shows that more than half of the experiments were performed in aqueous solvent (H_2O and/or D_2O) and a third in solvents with TFE concentration higher than 20%. The pH of most of the experiments is lower than 7 (Figure 5). All deposited experiments have proton chemical shift data, only one has carbon chemical shifts and 25 have amide proton chemical shift temperature coefficients. Data on secondary structure is scarce. Only 84 entries have information on global structure population and 25 entries have qualitative structured data.

Figure 6 compares the amino acid distribution in peptides of our database to that of proteins in SWISS-PROT (Bairoch and Apweiler, 2000). The two distributions are quite similar overall. An obvious exception is alanine, which occurs about twice as often in Pescador as in SWISS-PROT. This high proportion of alanine, known for its strong helical propensity, is related to the fact that a large proportion of the deposition entries represent peptides with sequences engineered to form helices used in deriving

Table 1. Pescador data content. The text in **bold** (first column) represents the different sections of the database. In the section '**experimental data**', the Table 'ChemicalShiftValues' contains the δ-values, 'CCValues' contains the coupling constant values, and the Table 'NH_TemperatureCoeff' contains the NH chemical shift temperature coefficients. The explanation of all the other ones can be found in the paragraph 'Database schema, and organizations'

Table name	Total number	Number of experiments
experiment		
Experiment_information	233	
experimental conditions		
pH [1, 5[		*103*
pH [5, 7[		*117*
pH [7, 10[		*12*
pH [10, 14]		*0*
Solvents	22	
aqueous solvent (H_2O and/or D_2O)		*160*
TFE solvent (> 20%)		*71*
AddComponents	19	86
global information		
Labs	15	
Refs	52	
Authors	119	
peptide data		
Peptides	145	
PeptideResidues	2461	
CisResidues	4	
SupplementaryCovBonds	2	
experimental data		
ChemicalShiftValues	15464	233
proton chemical shift data	*15442*	*232*
carbon chemical shift data	*22*	*1*
CCValues	10	1
NH_TemperatureCoeff	313	25
structure data		
Global_population	64	84
QualStructAnalValues	264	25

the Agadir prediction program (Muñoz and Serrano, 1994, 1995a,b; Lacroix et al., 1998). The bias introduced by these peptides is discussed in the following section. Another exception is cysteine, which occurs with a very low frequency for the Pescador set of small peptides. As more data becomes available, most of these differences are expected to disappear, although some residue biases due to preferences in peptide sequence engineering are likely to remain.

Definition of restricted data subset

With the aim of deriving a good set of sequence dependent 'random coil' δ-values for each residue type from the data deposited in Pescador, it is necessary to define a subset of the data that contains only peptides that exhibit little or no conformational preference in solution.

Within the full data set in Pescador, which comprises peptides with helical and β-sheet structure, as well as unstructured ones, we have defined a restricted subset that excludes peptides that were reported to adopt more than 20% overall secondary structure. We furthermore excluded from the analysis residues within peptides that were reported to be in a well defined conformation, peptides in solvents other than water, and residues at peptide termini. From the re-

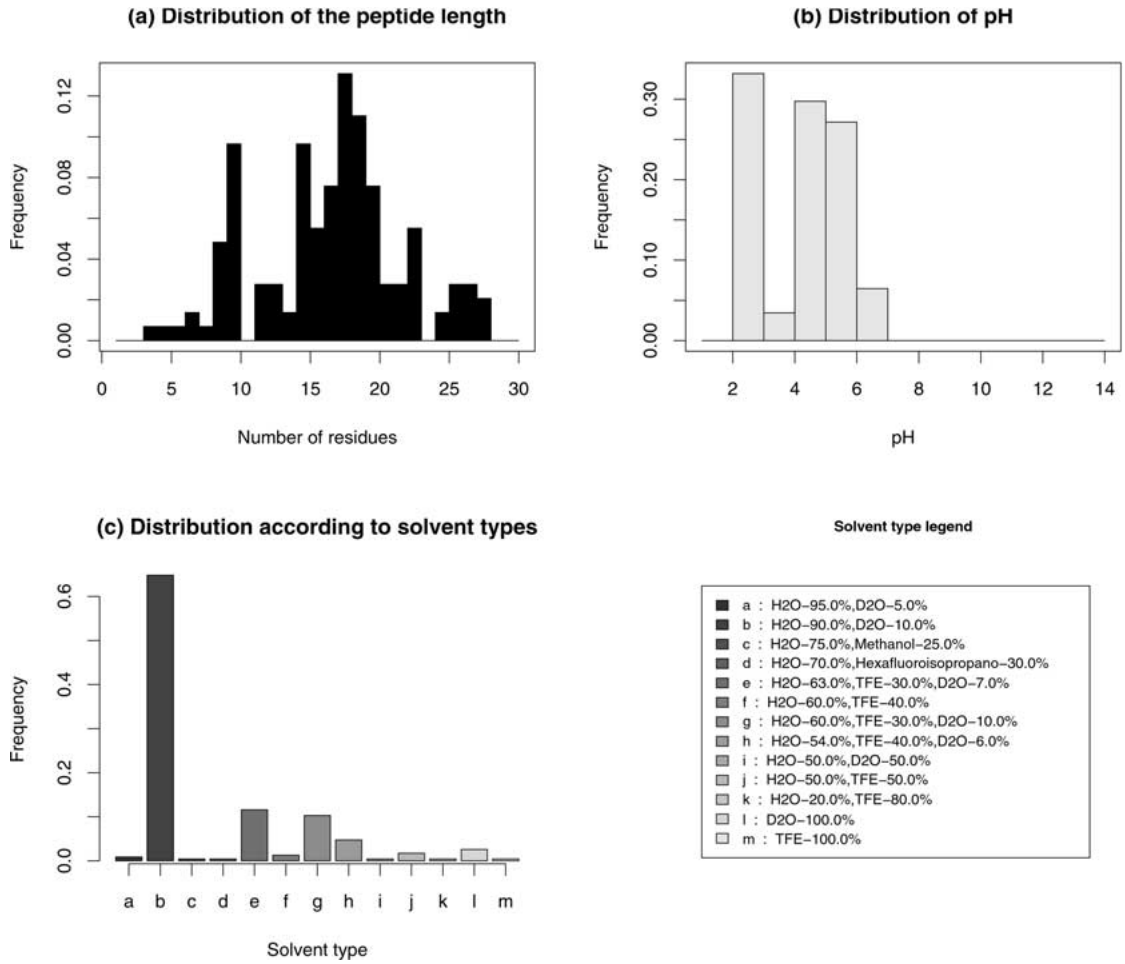

Figure 5. Pescador data distributions. (a) The distribution of peptide lengths, (b) the frequency of experiments as a function of pH, (c) frequency of experiments per solvent category.

sults displayed in Table 2, we see that the average alpha proton δ-values for this restricted set are similar, but generally larger (0.02 ppm, on average) than those of the full set. This likely reflects the presence in the full set of many helical peptides from the Agadir analysis. Some average δ-values display notable differences. Namely, the average δ-values for Ile and Met are respectively 0.08 ppm and 0.11 ppm higher in the restricted set, while the Trp value is 0.05 ppm lower. The interquartile range for the values is also smaller (on average 0.11 ppm in the restricted set compared to 0.16 ppm in the full set). Given that the residue subset selected here better represents conformationally averaged peptides, only this subset was used for the further analysis in this work.

The 'preferred random coil' alpha proton δ-values from Pescador

Typically, chemical shift resonance values for backbone nuclei in peptides and proteins are used to derive information on the conformational preferences of the backbone. These preferences are used in turn to describe the secondary structure adopted by the peptide or protein segment (Wishart et al., 1992). To be able to proceed in this manner, it is necessary to compare the δ-values measured for the peptide or protein at hand, with a set of well defined reference values, preferably corresponding to peptide or residues featuring completely unstructured or 'random coil' conformation. Such 'random coil' values are usually derived from series of small glycine-based model peptides (Bundi and Wüthrich, 1979; Merutka et al., 1995; Wishart et al., 1995). There are however, relatively large discrepan-

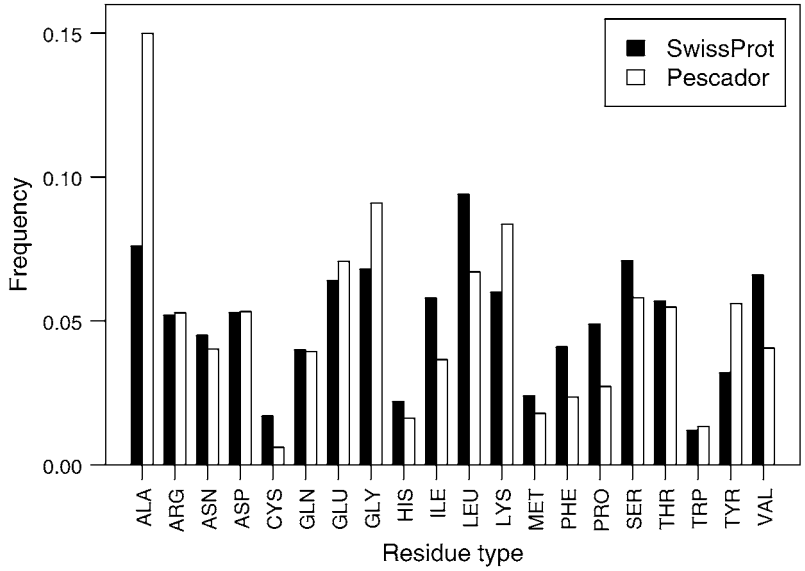

Figure 6. Residue frequencies in peptides deposited in Pescador (all peptides) (light) and in protein sequences in SWISS-PROT (dark) databases.

cies between the sets of 'random coil' δ-values derived by different authors. In the case of alpha protons, these differences can range from 0.03 up to 0.14 ppm for carboxyl-titrating residues (see relative graphs a–c, Figure 7). It has also been noted that using these references values alone is often not good enough, as there may be additional sequence dependent effects due to influences from neighboring residues. Recently a set of correction factors has been proposed to compensate for strong direct sequence effects (Schwarzinger et al., 2001), again based on a set of glycine-based peptides.

Table 2 gives an overview of the number of all alpha proton δ-values extracted from the restricted peptide set. We could establish using the Kolmogorov–Smirnov (Conover, 1971) and Shapiro–Wilk (Royston, 1982) tests for normality that none of the computed distributions were normal. It was therefore not possible to apply standard statistical tests on these data, and hence only a descriptive analysis based on mean and spread measures is presented to obtain an overview of the data. The absence of a normal distribution for the alpha proton δ-values is expected from the diversity of the peptides chemical shift data collected under different conditions, and subjected to many different factors.

Treatment and correction factors of alpha proton δ-values

Correction factors were calculated relative to the median of alpha proton δ-values for the 20 common amino acid types of the restricted set (Table 2, $\delta^{\text{ref.}}_{\text{restricted set}}$ column) derived from the database as reference. The correction factor A for an amino acid residue type z for the alpha proton chemical shift value was derived as follows:

$$A[z] = \frac{\sum_{i=1}^{20}\left(\sum_{j=1}^{n_i}\left(\delta X_{aj}[i] - \delta X_{\text{ref.}}[i]\right)\right)}{\sum_{i=1}^{20} n_i}, \qquad (1)$$

where $\delta X_{aj}[i]$ is the alpha proton δ-value of the residue X_a of type i in the subsequence

$$\ldots - X_a - X_b - z - X_c - X_d - \ldots \qquad (2)$$

for a specific z residue and $\delta X_{\text{ref.}}[i]$ is the median of alpha proton δ-values for the i amino acid type of restricted set (Table 2, $\delta^{\text{ref.}}_{\text{restricted set}}$ column). A summation is done over all the n_i possible subsequences and over all the twenty i amino acid types. The three other correction factors B, C and D were derived in the same way. Application of the correction factors to the Pescador mean shifts to compensate for sequence dependence was based on the formula by Schwarzinger et al. (2001):

Table 2. Alpha proton δ-values for the 20 common amino acids (number of amino acid (nr) and median ± interquartile range). Alpha proton δ-value differences (Δδ) are calculated as $\Delta\delta = \delta_{\text{complete set}} - \delta^{\text{ref.}}_{\text{restricted set}}$. Chemical shift differences bigger than 0.08 ppm are identified in **bold**

Residue	Complete set		Restricted set		Δδ
	No.	$\delta_{\text{complete set}}$	No.	$\delta^{\text{ref.}}_{\text{restricted set}}$	
ALA	537	4.24 ± 0.09	164	4.27 ± 0.08	−0.03
ARG	209	4.27 ± 0.17	77	4.30 ± 0.09	−0.03
ASN	168	4.70 ± 0.10	79	4.71 ± 0.11	−0.01
ASP	224	4.64 ± 0.13	96	4.63 ± 0.09	0.01
CYS	13	4.57 ± 0.15	9	4.58 ± 0.31	−0.01
GLN	154	4.28 ± 0.16	70	4.32 ± 0.09	−0.04
GLU	282	4.23 ± 0.15	90	4.28 ± 0.12	−0.05
GLY	504	3.98 ± 0.07	193	3.97 ± 0.06	0.01
HIS	53	4.70 ± 0.11	19	4.71 ± 0.03	−0.01
ILE	150	4.08 ± 0.29	57	4.16 ± 0.07	**−0.08**
LEU	291	4.28 ± 0.13	116	4.31 ± 0.09	−0.03
LYS	357	4.25 ± 0.16	135	4.28 ± 0.10	−0.03
MET	76	4.36 ± 0.25	19	4.47 ± 0.10	**−0.11**
PHE	92	4.59 ± 0.16	37	4.60 ± 0.08	−0.01
PRO	92	4.42 ± 0.10	38	4.42 ± 0.10	0.00
SER	203	4.41 ± 0.14	74	4.44 ± 0.12	−0.03
THR	225	4.33 ± 0.17	90	4.34 ± 0.09	−0.01
TRP	46	4.59 ± 0.42	24	4.54 ± 0.20	0.05
TYR	219	4.50 ± 0.17	62	4.54 ± 0.08	−0.04
VAL	163	4.08 ± 0.17	70	4.10 ± 0.10	−0.02
Total number	4058		1519		
Average		0.16		0.11	−0.02
Absolute average					0.03

$$\begin{aligned}\delta_R\text{corrected} &= \delta_R\text{ref.} + \Delta\delta_{R_{-1}} + \Delta\delta_{R_{+1}} \\ &\quad + \Delta\delta_{R_{-2}} + \Delta\delta_{R_{+2}} \\ &= \delta_R\text{ref.} + C_{[z=R_{-1}]} + B_{[z=R_{+1}]} \\ &\quad + D_{[z=R_{-2}]} + A_{[z=R_{+2}]}, \qquad (3)\end{aligned}$$

where the residue R is in the subsequence:

$$\ldots - R_{-2} - R_{-1} - R - R_{+1} - R_{+2} - \ldots. \qquad (4)$$

Effects of neighboring residues on alpha proton δ-values

The effect of neighboring residues on the 'default' δ-values for the amino acid backbone has long been recognized, especially for Pro preceding residues. Figure 8 shows the alpha proton δ-value distribution computed from data in Pescador for residues preceding four representative residues (Ala, Lys, Ile and Trp) compared to those preceding Pro. The clear difference between the distributions of the Pro set shows that the Pescador data reproduces this trend well.

A detailed analysis of neighboring effects was recently performed by Schwarzinger et al., who examined a series of GGXGG peptides in order to obtain sequence dependent correction factors for 'random coil' NMR chemical shifts (Schwarzinger et al., 2001). A similar analysis was performed here using chemical shift data stored in Pescador. To avoid interference from secondary structure formation, the restricted data set defined above was examined. The results given in Table 3 show that the significant downfield shift of the alpha proton chemical shift of residues preceding Pro is clearly present. Furthermore, strong influences of 0.08 ppm or more are observed for residues preceding His, Thr and Tyr and the $(i + 2)$ to $(i - 2)$ residues neighboring Trp. The trends observed for Cys are based on a limited body of data and are not commented.

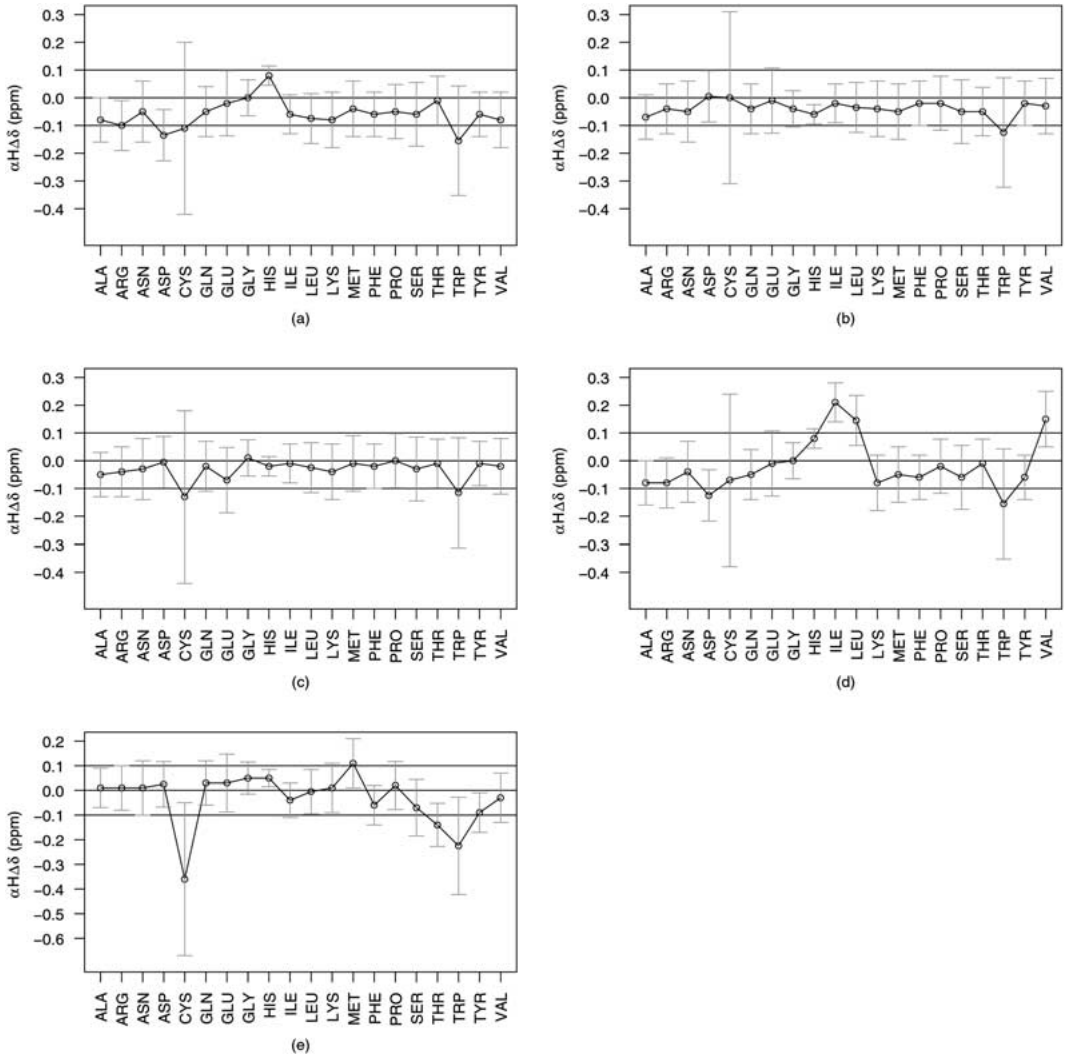

Figure 7. Alpha proton chemical shifts differences for the 20 common amino acids compared to 'random coil' values reported in literature. These differences are obtained by substracting the Pescador values of the restricted set from corresponding values from the series: (a) GGXA (308 K, pH 7.0, Bundi and Wüthrich (1979)), (b) GGXGG (between 278 K to 328 K, pH 5.0, Merutka et al. (1995)), (c) GGXAGG (298 K, pH 5.0, Wishart et al. (1995)), (d) Chemical Shift Index (Wishart et al., 1992), (e) BMRB mean values. The median (middle of the bar) and interquartile range (bar height represents the interval in which 50% of values reside) are included to give an idea of the Pescador data distribution.

The trends observed for Pro and Trp are similar, to those observed in the study of Schwarzinger et al. but more pronounced. The differences observed for the residues preceding Thr and His are also larger in Pescador (0.08 ppm and −0.08 ppm), while not reaching the significance threshold for Schwarzinger et al. The differences observed for residues preceding and following Phe and Tyr are very similar in the Pescador set. The observed differences can be attributed to the fact that the Pescador data represents diverse flanking sequences, while the GGXGG set is biased by the presence of the Gly residues. A possible explanation for the discrepancy in the results for Phe, Tyr and Trp aromatic residues may be the absence (or presence) of aromatic side-chain – backbone interactions in the GGXGG series compared to the peptides in Pescador. In fact, the existence of Tyr(i)-Gly($i+2$) interactions have been demonstrated (Kemmink and Creighton, 1995; Kemmink et al., 1993). The larger deviation for residues preceding Pro and Thr can be related to the increased steric interactions expected for residues with a side chain compared to Gly.

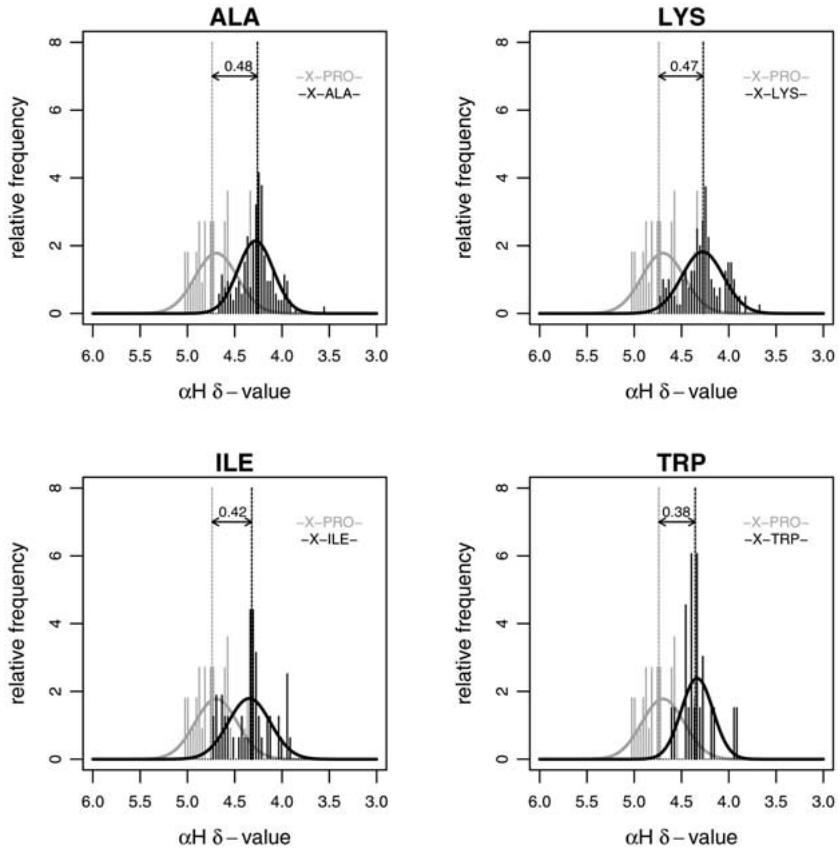

Figure 8. Comparison of the distribution of the alpha proton δ-value of residue X followed by the Pro (light) and the one followed by the residues Ala, Lys, Ile or Trp (dark) from the restricted set. The relative frequency of chemical shift is represented with vertical bars, the normal distribution curve is the fitted curve on the observed chemical shift frequency distribution and the dotted vertical line represents the median of each distibution. The difference between the two medians is above the arrow.

We further examined the average alpha proton δ-values of residues preceding Pro in the restricted set. The deviations from the mean Pescador values corrected by the *B* factor for Pro are between −0.09 and 0.03 ppm, depending on residue type. Based on the 'random coil' values for the GGXGG series (Merutka et al., 1995) combined with the Schwarzinger et al. correction factors (Schwarzinger et al., 2001) these differences are between 0.01 and 0.17 ppm.

Application of the correction factors

We tested the correction factors on a small sample of peptides that are not present in the database, the β-sheet forming Carp Granulin 1–30 (Vranken et al., 1999), a V3 Loop Fragment exhibiting some nascent helix (Chandrasekhar et al., 1991), and the internalization signal of Lysosomal Acid Phosphatase which forms a β-turn (Eberle et al., 1991). The peptide sequences are given in the legend of Table 4. The correction factors (Table 3) were applied to the Pescador median δ-values of the restricted set for each residue type in the simplest way, calculated according to Equation 3. They were considered to be additive, and all contributions were included (also the ones below 0.08 ppm). In all three test cases the absolute difference of the observed δ-values relative to the values calculated by Pescador were smaller than those computed relative to a standard set of 'random coil' values (Merutka et al., 1995), even after applying the correction factors (Schwarzinger et al., 2001) (Table 4). It is interesting to note that the corrected 'random coil' values (Table 4, column 3) give overall values halfway between the original 'random coil' values and the Pescador corrected values. The differences between the sets are also the most striking for the Lysosomal Acid Phosphatase, which has no conformational preference except for a central β-turn. For this particular case the alpha proton chemical shift

Table 3. Alpha proton chemical shifts correction factors for the 20 common amino acids in the subsequence $\ldots - X_a - X_b - z - X_c - X_d - \ldots$, calculated according to Equation 1 on the restricted set. Sequence dependent correction factors higger than 0.08 ppm are identified in **bold**. 'Ave.' indicates the average difference: the top row is the overall difference, the bottom row the absolute difference. n is the subsequence number per residue type z

z	A	B	C	D	n_i
ALA	-0.01 ± 0.10	-0.04 ± 0.09	-0.03 ± 0.09	-0.02 ± 0.08	109
ARG	0.00 ± 0.07	0.00 ± 0.06	0.00 ± 0.07	-0.01 ± 0.06	48
ASN	0.03 ± 0.08	0.00 ± 0.06	-0.01 ± 0.06	0.00 ± 0.07	45
ASP	0.00 ± 0.08	-0.01 ± 0.07	-0.02 ± 0.09	-0.02 ± 0.10	53
CYS	0.02 ± 0.02	0.00 ± 0.03	$\mathbf{0.21} \pm 0.11$	$\mathbf{-0.09} \pm 0.03$	3
GLN	0.01 ± 0.07	-0.03 ± 0.07	-0.01 ± 0.07	0.00 ± 0.09	37
GLU	-0.02 ± 0.11	-0.04 ± 0.07	-0.04 ± 0.09	0.01 ± 0.12	62
GLY	0.00 ± 0.08	0.00 ± 0.05	0.05 ± 0.10	0.01 ± 0.03	11
HIS	0.02 ± 0.21	$\mathbf{-0.08} \pm 0.03$	0.06 ± 0.07	-0.01 ± 0.02	8
ILE	0.00 ± 0.06	0.01 ± 0.06	0.02 ± 0.09	0.00 ± 0.06	36
LEU	-0.02 ± 0.12	-0.03 ± 0.08	-0.01 ± 0.07	-0.01 ± 0.07	78
LYS	-0.02 ± 0.09	-0.03 ± 0.07	-0.02 ± 0.06	0.01 ± 0.10	71
MET	-0.05 ± 0.05	-0.01 ± 0.08	0.01 ± 0.07	0.03 ± 0.13	12
PHE	-0.03 ± 0.06	-0.06 ± 0.08	-0.04 ± 0.19	-0.07 ± 0.08	24
PRO	0.04 ± 0.13	$\mathbf{0.30} \pm 0.06$	0.00 ± 0.13	0.04 ± 0.06	20
SER	0.01 ± 0.07	0.05 ± 0.04	0.03 ± 0.05	-0.02 ± 0.10	33
THR	0.02 ± 0.05	$\mathbf{0.08} \pm 0.04$	0.03 ± 0.05	0.01 ± 0.11	52
TRP	$\mathbf{-0.08} \pm 0.16$	$\mathbf{-0.16} \pm 0.11$	$\mathbf{-0.16} \pm 0.08$	$\mathbf{-0.16} \pm 0.09$	15
TYR	-0.02 ± 0.07	$\mathbf{-0.08} \pm 0.11$	-0.06 ± 0.08	-0.01 ± 0.06	32
VAL	-0.01 ± 0.12	0.02 ± 0.12	0.03 ± 0.07	0.00 ± 0.09	41
Ave.	-0.01 ± 0.09	0.00 ± 0.07	0.00 ± 0.08	-0.02 ± 0.08	
	0.02 ± 0.09	0.05 ± 0.07	0.04 ± 0.08	0.03 ± 0.08	

difference for the Gln4, Pro5 and Pro6 residues is (0.22, 0.26 and −0.03 ppm) compared to Merutka et al. (1995), (0.11, 0.15 and −0.03 ppm) compared to Merutka et al. (1995) with correction factors of Schwarzinger et al., and (−0.05, 0.01 and 0.01 ppm) for the values calculated from Pescador.

These results show that the protocol currently employed by Pescador to calculate the sequence dependent shift values yields similar results to those obtained using values derived from glycine-based peptides, but also that some interesting differences are present. Clearly, the current protocol is open for improvement, as pure additivity of the correction factors implies that steric and/or field-related effects of residues on their neighbors are independent of each other. Furthermore, additional data from peptides exhibiting no conformational preferences in solution will greatly enhance the accuracy of the mean values calculated for the restricted subset. This will also allow a more thorough investigation of the best way to implement the correction factors. Overall, the analysis of diverse peptide sequences appears to present a more realistic view of peptide conformation in solution, and should therefore produce more reliable correction factors. Given a large enough data set, these correction factors can even be residue-specific: for example, a different correction factor could be derived for a Gly residue preceding a Pro as compared to other residue pair combinations.

Effects of TFE on amide proton δ*-values*

The addition of TFE at a given temperature results in a decrease of amide proton δ-values (Merutka et al., 1995). This is confirmed for selected amino acids, such as Ile, Leu, Asp, Thr, Arg and Trp in peptides stored in Pescador that have been studied at different TFE concentration at 278 K (Figure 9). Although the number of data points is limited in each case, the de-

Table 4. Global average deviations of experimental alpha proton δ-values for the Carp Granulin 1–30 peptide (Vranken et al., 1999), a V3 Loop Fragment (Chandrasekhar et al., 1991), and the Lysosomal Acid Phosphatase (Eberle et al., 1991). The absolute deviation (AD), only the negative contributions (NEG), and only the positive contributions (POS) are shown against the expected δ-values from Pescador (column 1), from Merutka et al. (column 2, 1995), and from Merutka et al. with correction factors (column 3, Schwarzinger et al. (2001))

	Pescador			Merutka			Merutka + Schwarzinger		
Peptide	AD	NEG	POS	AD	NEG	POS	AD	NEG	POS
Carp granulin 1–30[a]	0.408	−0.217	0.191	0.426	−0.222	0.203	0.429	−0.213	0.215
V3 loop fragment[b]	0.067	−0.036	0.031	0.075	−0.053	0.021	0.071	−0.045	0.026
Lysosomal acid phos.[c]	0.069	−0.049	0.020	0.111	−0.082	0.029	0.089	−0.072	0.017

[a] VIHCDAATICPDGTTCSLSPYGVWYCSPFS.
[b] YNKRKRIHIGPGRAFYTTKNIIGC.
[c] MQAQPPGYRHVADGEDHA.

crease of amide proton chemical shift continues up to 80% TFE (Leu, Asp and Thr graphs). The observed decrease in amide proton δ-values most probably reflects a parallel decrease in amide-water hydrogen bonding. A similar but smaller variation was observed for the alpha proton δ-values as a function of TFE. In this case, the negative Δδ values must be probably due to an increase in helical conformations. For example, the alpha proton δ-value of Leu moves 0.10 ppm upfield on average when the TFE concentration is increased from 10% to 80%, whereas its amide proton moves 0.54 ppm towards lower δ-values on average.

The confirmation of the trend observed by Merutka et al. (1995), even with the limited data set in Pescador, clearly illustrates the potential for further analysis on the effects of TFE on amide proton chemical shift and secondary structure, once sufficient data become available.

Discussion

The importance of gathering data on peptides with varying populations of conformations has already been illustrated with the work on Agadir, a program for predicting α-helical content of peptides (Muñoz and Serrano, 1994, 1995a,b; Lacroix et al., 1998). With Pescador, data capture is extended to a wide range of peptides with different conformational preferences. In this respect, one of the main advantages of Pescador is that the effects of many different parameters, such as temperature and pH, can be analyzed and predicted from the experimentally observed characteristics of amino acids without requiring the actual experimental analysis of the series of corresponding peptides. Furthermore, the primary experimental data on peptide conformation are derived from a large number of sources, and analysis on these data reduces the influence of laboratory-specific approaches and should help in drawing more reliable conclusions. Also, less assumptions about peptide behavior are required in comparison to examining a limited set of peptides. This versatility and diversity should make Pescador an excellent tool for the identification of peptide secondary structure and its relation to amino acid sequence, as well as identifying potentially interesting conformational effects which require more detailed experimental validation. The limited amount of data in Pescador already enables to reproduce values and trends reported in literature. This suggests that Pescador is likely to become even more useful in this regard as the amount of deposited data increases in the future.

Although the analysis presented here is focussed on the alpha proton δ-values, it is directly transferable to other atom types which are often more reliable for secondary structure determination. However, heteronuclear spectra are usually not recorded for peptides and a more reliable method for determining secondary structure based on the alpha proton remains extremely useful for work on peptide conformation. In a conformationally averaged peptide, the alpha proton δ-value of a nucleus is mainly defined by the (averaged) value of the phi backbone angle. The preferred phi value for a residue in such a peptide is expected to be determined firstly by the type of side chain of that residue, which might interact with the backbone and impose steric constraints on its conformation, and secondly by possible effects of neighboring residues. Two different residues with the same preferred phi value will, however, not necessarily have the same alpha proton δ-value, since the residue side chain and interactions with neighboring

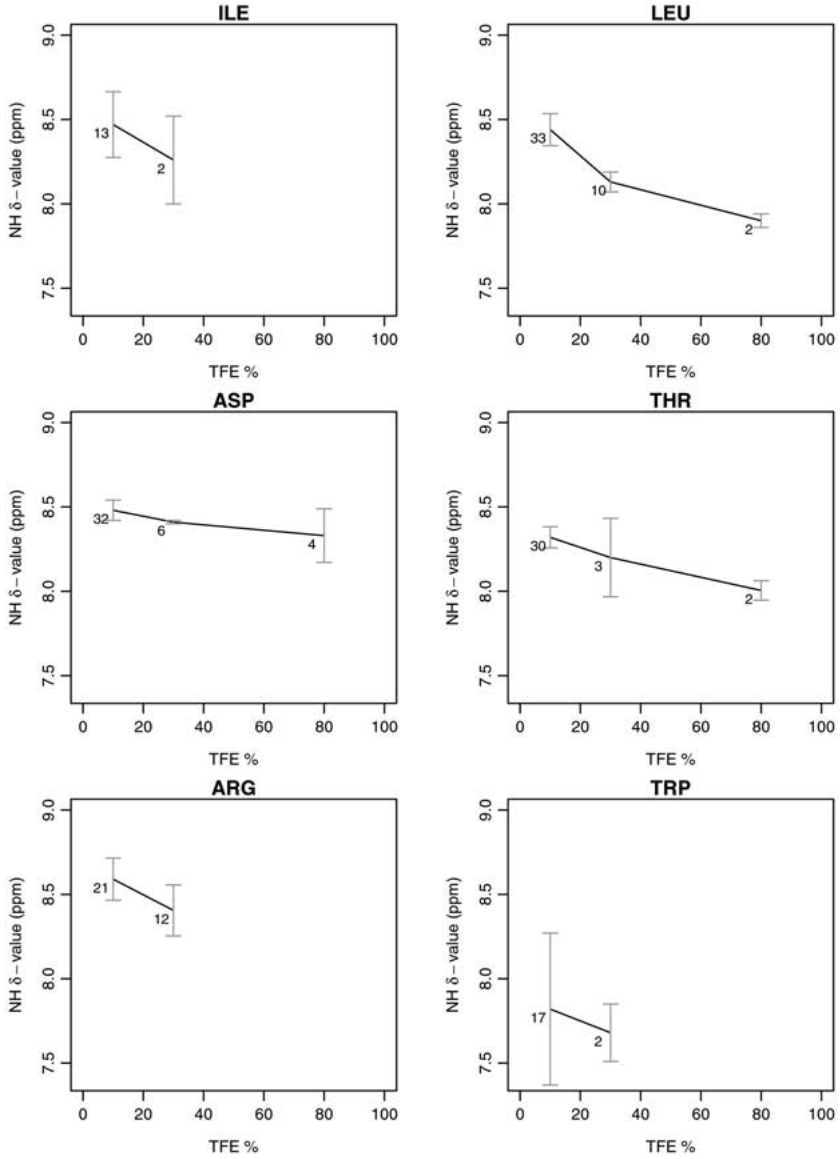

Figure 9. Evolution of the average amide proton δ-value of TFE restricted set of selected amino acids with increasing TFE concentration at 278K. Light bar height represents the interval in which 50% of values are and the numbers at next are the data number for each point.

residues will have a different effect on the chemical environment of the alpha proton. In this respect, the Pescador data provides sequence dependent alpha proton δ-values for a given residue type in its most 'preferred random coil' conformation in solution when it is part of a larger peptide that adopts no obvious secondary structure. These 'preferred random coil' values are inherently different from the chemical shift index values, which are obtained from proteins with known secondary structure, and from the 'random coil' values, which are obtained from series of short glycine-based peptides. As such, the 'preferred random coil' values also provide a good starting point for assessing secondary structure formation in peptides. Since the preferred phi dihedral angle will be different for each residue type, the size of the deviations from this value that indicate secondary structure formation are also residue-specific. Given a large enough database with detailed information on peptides adopting secondary structure, it should be possible to calibrate the deviations from the 'preferred random coil' values with respect to secondary structure content.

Pescador currently holds primarily NMR data. The importance of CD data for the database is clear, it is a very commonly used method for obtaining global secondary structure preferences for peptides. This is especially interesting for the peptides examined by NMR, as their experimental parameters on a residue basis can be correlated with their global secondary structure content.

Conclusion

Pescador is a database with a single focus on conformational peptide data. The reduced data model that is allowed by this approach does limit the scope of data that can be handled, but on the other hand it allows a simple and quick deposition system as well as easier data processing. Furthermore, the relational database linked to the depositions enables complex analysis of the data and easy accessibility by outside users. This database offers a novel means of obtaining expected values for NMR observable based on peptide or protein sequence and environment. Most importantly, bias is reduced, as it allows the examination of many different peptides in many different conditions from a wide range of laboratories. The availability of the information in a well formatted relational database allows for easy and extensive validation and analysis of the data. The potential for a wide range of searches on the influence of, for example, sequence, pH and temperature effects on especially NMR parameters is present. However, the bottleneck for improved analyses remains the curation of data, and we would therefore like to end with an appeal to all groups involved in peptide research to deposit their data in Pescador.

Acknowledgements

We are grateful to Jean Richelle and Christian Lemer for valuable help in setting up and analyzing the database. We would like to thank Luis de la Vega for data deposition, BMRB for reference and initial discussions, and Luis Serrano for providing the Agadir peptide data. Pescador has been developed in the framework of the research project entitled: 'De-Novo Design of Peptides with Well Defined Conformation and Function', supported by the European Commission, Contract number: BIO4 CT 972086.

References

Andersen, N.H. and Tong, H. (1997) *Protein Sci.*, **6**, 1920–1936.
Bairoch, A. and Apweiler, R. (2000) *Nucl. Acids Res.*, **28**, 45–48.
Baldwin, R.L. (1995) *Biophys. Chem.*, **55**, 127–135.
Bundi, A. and Wüthrich, K. (1979) *Biopolymers*, **18**, 285–298.
Chakrabartty, A., Schellman, J.A. and Baldwin, R.L. (1991) *Nature*, **351**, 586–588.
Chandrasekhar, K., Profy, A.T. and Dyson, H.J. (1991) *Biochemistry*, **30**, 9187–9194.
Conover, W.J. (1971) *Practical Nonparametric Statistics*, John Wiley and Sons, New York, NY.
Eberle, W., Sander, C., Klaus, W., Schmidt, B., von Figura, K. and Peters, C. (1991) *Cell*, **67**, 1203–1209.
Gellman, S.H. (1998) *Curr. Opin. Chem. Biol.*, **2**, 717–725.
Griffiths-Jones, S.R. and Searle, M.S. (2000) *J. Am. Chem. Soc.*, **122**, 8350–8356.
Griffiths-Jones, S.R., Maynard, A.J. and Searle, M.S. (1999) *J. Mol. Biol.*, **292**, 1051–1069.
Griffiths-Jones, S.R., Sharman, G.J., Maynard, A.J. and Searle, M.S. (1998) *J. Mol. Biol.*, **284**, 1597–1609.
Ihaka, R. and Gentleman, R. (1996) *J. Comput. Graphical Stat.*, **5**, 299–314.
Kemmink, J. and Creighton, T.E. (1995) *J. Mol. Biol.*, **245**, 251–260.
Kemmink, J., van Mierlo, C.P., Scheek, R.M. and Creighton, T.E. (1993) *J. Mol. Biol.*, **230**, 312–322.
Lacroix, E., Kortemme, T., de la Paz, M.L. and Serrano, L. (1999) *Curr. Opin. Struct. Biol.*, **9**, 487–493.
Lacroix, E., Viguera, A.R. and Serrano, L. (1998) *J. Mol. Biol.*, **284**, 173–191.
Lopez-Hernandez, E. and Serrano, L. (1995) *Fold Des.*, **1**, 43–55.
Lopez-Hernandez, E. and Serrano, L. (1996) *Fold Des.* **1**, 43–55.
Markley, J.L., Bax, A., Arata, Y., Hilbers, C.W., Kaptein, R., Sykes, B.D., Wright, P.E. and Wüthrich, K. (1998) *J. Mol. Biol.*, **280**, 933–952.
Merutka, G., Dyson, H.J. and Wright, P.E. (1995) *J. Biomol. NMR*, **5**, 14–24.
Millhauser, G.L., Stenland, C.J., Bolin, K.A. and van de Ven, F.J. (1996) *J. Biomol. NMR*, **7**, 331–334.
Muñoz, V. and Serrano, L. (1994) *Nat. Struct. Biol.*, **1**, 399–409.
Muñoz, V. and Serrano, L. (1995a) *J. Mol. Biol.*, **245**, 275–296.
Muñoz, V. and Serrano, L. (1995b) *J. Mol. Biol.*, **245**, 297–308.
Muñoz, V., Serrano, L., Jimenez, M.A. and Rico, M. (1995) *J. Mol. Biol.*, **247**, 648–669.
Odaert, B., Jean, F., Boutillon, C., Buisine, E., Melnyk, O., Tartar, A. and Lippens, G. (1999) *Protein Sci.*, **8**, 2773–2783.
Padmanabhan, S. and Baldwin, R.L. (1994) *Protein Sci.*, **3**, 1992–1997.
Royston, P. (1982) *Appl. Stat.*, **31**, 176–180.
Santiveri, C.M., Rico, M. and Jimenez, M.A. (2000) *Protein Sci.*, **9**, 2151–2160.
Santiveri, C.M., Rico, M. and Jimenez, M.A. (2001) *J. Biomol. NMR*, **19**, 331–345.
Schwarzinger, S., Kroon, G.J., Foss, T.R., Chung, J., Wright, P.E. and Dyson, H.J. (2001) *J. Am. Chem. Soc.*, **123**, 2970–2978.
Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J.L. (1991) *J. Biomol. NMR*, **1**, 217–236.
Taddei, N., Chiti, F., Fiaschi, T., Bucciantini, M., Capanni, C., Stefani, M., Serrano, L., Dobson, C.M. and Ramponi, G. (2000) *J. Mol. Biol.*, **300**, 633–647.

Vranken, W.F., Chen, Z.G., Xu, P., James, S., Bennett, H.P. and Ni, F. (1999) *J. Pept. Res.*, **53**, 590–597.
Wishart, D.S., Bigam, C.G., Holm, A., Hodges, R.S. and Sykes, B.D. (1995) *J. Biomol. NMR*, **5**, 67–81.
Wishart, D.S., Sykes, B.D. and Richards, F.M. (1992) *Biochemistry*, **31**, 1647–1651.
Zerella, R., Chen, P.Y., Evans, P.A., Raine, A. and Williams, D.H. (2000) *Protein Sci.*, **9**, 2142–2150.