Journal of Biomolecular NMR, **12:** 395–405, 1998.
KLUWER/ESCOM

CAMRA: Chemical shift based computer aided protein NMR assignments

Wolfram Gronwald[a,b], Leigh Willard[a], Timothy Jellard[a], Robert F. Boyko[b], Krishna Rajarathnam[a,b], David S. Wishart[a,c], Frank D. Sönnichsen[d] and Brian D. Sykes[a,b,*]
[a]*Protein Engineering Network of Centres of Excellence, 713 Heritage Medical Research Centre, University of Alberta, Edmonton, AB, Canada T6G 2S2;* [b]*Department of Biochemistry, University of Alberta, Edmonton, AB, Canada T6G 2H7;* [c]*Faculty of Pharmacy and Pharmaceutical Sciences, University of Alberta, Edmonton, AB, Canada T6G 2N8;* [d]*Department of Physiology and Biophysics, Case Western Reserve University, Cleveland, OH 44106-4970, U.S.A.*

Received 26 March 1998; Accepted 8 June 1998

Key words: assignment homology programs, automatic

Abstract

A suite of programs called CAMRA (Computer Aided Magnetic Resonance Assignment) has been developed for computer assisted residue-specific assignments of proteins. CAMRA consists of three units: ORB, CAPTURE and PROCESS. ORB predicts NMR chemical shifts for unassigned proteins using a chemical shift database of previously assigned homologous proteins supplemented by a statistically derived chemical shift database in which the shifts are categorized according to their residue, atom and secondary structure type. CAPTURE generates a list of valid peaks from NMR spectra by filtering out noise peaks and other artifacts and then separating the derived peak list into distinct spin systems. PROCESS combines the chemical shift predictions from ORB with the spin systems identified by CAPTURE to obtain residue specific assignments. PROCESS ranks the top choices for an assignment along with scores and confidence values. In contrast to other auto-assignment programs, CAMRA does not use any connectivity information but instead is based solely on matching predicted shifts with observed spin systems. As such, CAMRA represents a new and unique approach for the assignment of protein NMR spectra. CAMRA will be particularly useful in conjunction with other assignment methods and under special circumstances, such as the assignment of flexible regions in proteins where sufficient NOE information is generally not available. CAMRA was tested on two medium-sized proteins belonging to the chemokine family. It was found to be effective in predicting the assignment providing a database of previously assigned proteins with at least 30% sequence identity is available. CAMRA is versatile and can be used to include and evaluate heteronuclear and three-dimensional experiments.

Abbreviations: CAMRA, Computer Aided Magnetic Resonance Assignment; GUI, graphical user interface; IL-8, interleukin-8; NOE, nuclear Overhauser effect; SDF-1, stromal derived factor-1; IPP, interactive peak picker; SSS, spin system separation.

Introduction

Nuclear magnetic resonance spectroscopy is widely used for the determination of protein structures in solution. However, the time required to solve a new protein NMR structure can vary from months to years, with the residue-specific assignment of the NMR spectra being the time limiting step. For small or medium-sized proteins the sequential assignment process relies primarily on conventional two-dimensional methods (Wüthrich, 1986). For large proteins, several strategies have been proposed for sequence-specific assignments, based on the combination of various heteronuclear experiments using ^{13}C and ^{15}N labeled proteins (for reviews see Clore and Gronenborn, 1991 and Bax

*To whom correspondence should be addressed.

and Grzesiek, 1993). One obvious approach to make the assignment process faster is to fully or partly automate it. Over the last few years, several computer methods have been proposed to accomplish this task (Kleywegt et al., 1991; Oschkinat et al., 1991; Xu and Sanctuary, 1993; Meadows et al., 1994; Hare and Prestegard, 1994; Kjaer et al., 1994; Kraulis, 1994; Friedrichs et al., 1994; Zimmermann et al., 1994; Olson and Markley, 1994; Morelle et al., 1995; Xu et al., 1995; Bartels et al., 1996; Lukin et al., 1997; Croft et al., 1997). With the exception of 3D/4D heteronuclear experiments using labeled proteins, all of the above automated assignment procedures require the use of NOEs to obtain residue-specific NMR assignments.

In this paper, we describe a suite of programs that has been developed to enable NMR spectroscopists to obtain residue-specific assignments for proteins without any sequential connectivity data. CAMRA (Computer Aided Magnetic Resonance Assignment) uses predicted chemical shift information and computer-identified spin systems to obtain residue-specific assignments for the protein of interest. For this task, three independent computer programs (ORB, CAPTURE and PROCESS) have been designed and combined into one package. ORB, which has been described previously (Gronwald et al., 1997), uses sequence and chemical shift similarity to predict the chemical shifts for the protein of interest. The goal of CAPTURE is two-fold: (1) to help the NMR spectroscopist generate a list of valid peaks from a two dimensional NMR spectrum, by filtering out noise peaks and other artifacts and (2) to separate the obtained peak list into distinct spin systems for assignment by other programs. PROCESS uses statistical weighting functions to combine the chemical shift predictions obtained by ORB with the spin systems identified by CAPTURE to propose assignments for the protein of interest. PROCESS offers several ranked choices for each assignment so that the user can confirm these assignments with additional spectral information. One main advantage of CAMRA is that the user has to check only two or three choices for each assignment as opposed to analyzing the entire spectrum.

The CAMRA suite of programs is especially useful when a series of homologous proteins are studied. An example would be a series of mutants studied in the same laboratory under similar conditions in which case the differences in buffer conditions, pH, temperature, and instrumental variations are minimized. In the process of assigning a series of homologous proteins, the information content of each newly assigned protein can be added to the user-supplied database of homologous proteins to assist with the assignment of related proteins. Several programs have been reported in the literature that use NMR chemical shifts of previously assigned proteins (Hare and Prestegard, 1994; Bartels et al., 1996). However, to the best of our knowledge, CAMRA is the first package to use a multi-protein database of previously assigned homologues to obtain a residue-specific assignment. In contrast to other packages, CAMRA does not use any connectivity information between the various spin systems to obtain a residue-specific assignment.

In assessing the strengths and limitations of this approach, CAMRA was tested on two members of the CXC chemokine family: an interleukin-8 analog, which contains a single point mutation in comparison to the wild type, and the stromal derived factor-1 (SDF-1). More than fifty chemokine sequences are known from human, other mammals and viruses (Baggiolini et al., 1996). Eight chemokine structures have been solved by NMR spectroscopy (Fairbrother and Skelton, 1996) and all of them adopt a similar tertiary fold of three β strands and an overlying α helix. Our group is interested in understanding the structure-function relationship in chemokines, in particular IL-8, for their clinical relevance (Rajarathnam et al., 1994, 1995; Crump et al., 1997). We are using NMR as a tool to study the structure-activity relationship and are actively exploring the feasibility of designing better agonists and antagonists for IL-8.

To predict chemical shifts for the IL-8 analog a database consisting of 8 previously assigned homologues was used. The database contains proteins with a very high degree of sequence similarity ($\geq$95%). For SDF-1 a database of 7 weakly homologous previously assigned proteins (18–30% sequence similarity) was available. For both proteins 2D TOCSY spectra were used to obtain peak lists and identify spin systems. From these results, the strengths and limitations of similarity based residue-specific assignments as used in CAMRA are discussed.

Methods

CAMRA consists of three parts: ORB, PROCESS, and CAPTURE (Figure 1). A description of these programs is presented here. Complete algorithmic details are available from the authors.

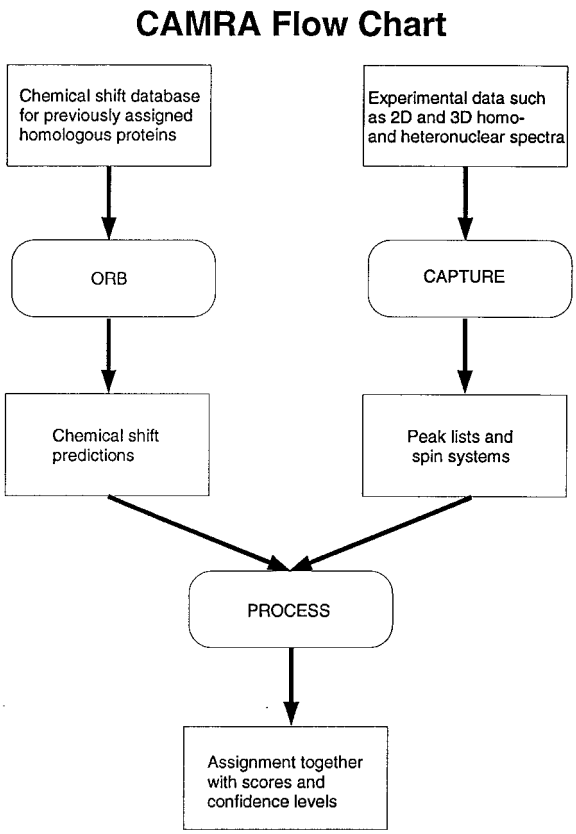

Figure 1. Flow chart describing the CAMRA suite of programs. All steps are further described in the algorithm section.

Program usage and programming details

The CAMRA suite of programs is designed to be user friendly. Whenever possible the CAMRA programs have easy-to-use graphical user interfaces (GUIs). All programs feature on-line documentation and have users' manuals. The peak picking section of CAPTURE, called IPP, is designed to be run from VNMR, which is the acquisition/processing software on all Varian spectrometers (Varian Associates Inc., Palo Alto, CA). It is the only part in the CAMRA package that is dependent upon specific software (however, other peak picking software may be used in place of IPP). There are a number of file conversion programs which are included with CAMRA. A detailed introduction into the usage of CAPTURE can be found in Bigam et al. (1997).

Most of the programming was done using the C programming language on the UNIX operating system. To create the GUIs, the Tk/Tcl programming package (Ousterhout, 1994) was used. In addition, PERL (Wall et al., 1990) and shell scripts were used to write utility programs. For the IPP program, VNMR/MagicalII macros (Varian Associates Inc., Palo Alto, CA) were used.

Description of ORB

ORB is used to predict the chemical shift of the query protein using a chemical shift database of previously assigned homologous proteins. For a complete description of ORB, see the previously published article (Gronwald et al., 1997).

Description of CAPTURE

The goal of CAPTURE is twofold: (1) to help the NMR spectroscopist generate a list of valid peaks from a two dimensional NMR spectrum, by filtering out noise peaks and other artifacts and (2) to separate the obtained peak list into distinct spin systems for assignment by other programs.

Step 1 – peak picking

IPP (interactive peak picker) is a set of programs for performing some basic operations on VNMR data, with the purpose of obtaining a list of valid peaks. IPP is a collection of C programs and VNMR/MagicalII macros that are accessible from the VNMR menu system or from the VNMR command line interface. IPP in itself does not pick peaks. The idea is that the spectroscopist is better at making peak-picking decisions than the computer. However, in order for the spectroscopist to make these decisions wisely, additional information is required. IPP provides this information by means of tools that perform tasks to correct the diagonal peaks, evaluate asymmetric peaks, add artificial symmetric peaks, and delete peaks.

Step 2 – separation of spin systems

SSS (spin system separation) is a program that decomposes a set of peaks from TOCSY spectra into sets of peak families or spin systems. The algorithm works in a number of steps (Figure 2):

- First gather a list of all non-diagonal peaks. From this, generate a list of main-diagonal peaks.
- Create a set of 'edges' from the list of peaks, by joining any peaks which are parallel on the same axis (discrimination values specified in a configuration file are used to asses if two peaks are parallel).
- Generate a set of 'boxes' by combining any four matching edges. The 'main diagonal boxes' is a

subset of 'boxes' that contains two main diagonal peaks.
- For each 'main diagonal box', recursively gather all boxes which match along one edge (Figure 2A). Add any 'main diagonal box' which has two edges in this set (Figure 2B). Then repeat this procedure starting with any 'main diagonal box' which was added, until there are no more main diagonal boxes which can be added (Figure 2C/D). These extracted boxes form one complete spin system (Figure 2D).

SSS then does some post-processing, to combine all remaining peaks of the fingerprint region of the spectrum into separate spin systems. In this stage, all peaks that are parallel on one axis are assigned to one spin system. After the computer automated generation of spin systems has been completed, the user may display and edit any one spin system superimposed on the spectrum. Note that because a peak can be in more than one box, it can be in more than one spin system. If peaks are missing in the spectrum, it is not possible to form and combine all necessary boxes for all spin systems. This will lead to truncated spin systems and also to peaks that have not been assigned to any of the assembled spin systems. In some cases the program is able to detect peaks missing from the spectra, and is robust enough to deal with some cases of complete peak overlap by allowing a single peak to be present in more than one spin system. Problems occur when the program encounters streaks where there are no real protein peaks across an entire band of the spectra (e.g. water streaks), and when one spin system is nearly (or completely) obscured by another. However, SSS performs well even when the amide peaks of two spin systems align along a single frequency.

We have found that SSS works best for medium-sized proteins which have a good spectral dispersion and which show a good TOCSY transfer. The best discriminated spin systems are the ones which do not overlap with any other spin systems, and have no missing peaks. Ala, Val, Asp and Asn are easy to discriminate because usually all peaks in these short spin systems are visible and do not overlap with each other or other spin systems. Long spin systems like Lys and Arg can cause problems because often peaks are missing or the peaks occur in a crowded region of the spectrum. Glycines can be difficult to identify due to the lack of side chain protons.

Description of PROCESS

The decision making part in the CAMRA package is called PROCESS. The goal of PROCESS is to obtain a possible residue-specific assignment for the protein of interest. To accomplish this, PROCESS calculates which group of observed NMR peaks (spin system) most likely correspond to which amino acid for the given protein. A central feature of PROCESS is that it does not need, nor does it use, any kind of sequential connectivity information (i.e. NOEs). Instead, residue-specific assignments are accomplished within PROCESS by matching a set of predicted shifts to a set of observed peaks. The predicted shifts can be obtained from a program such as ORB, or if available they can be calculated from an X-ray structure (de Dios et al., 1993), or by some other method of the users choosing. The observed peaks can be obtained from a variety of spectra such as 2D/3D TOCSY, 2D/3D DQF-COSY and 2D HSQC. It is important that all peaks which belong to one spin system are grouped together. Programs such as CAPTURE can be used to create these spin systems. PROCESS proposes assignments for each observed peak and spin system and performs a complete residue-specific assignment for the protein of interest. The following outlines the steps that PROCESS takes to arrive at its final output:

Step 1 – generating expected peaks

In general, only a subset of all theoretically possible NMR signals is observed. However, a program such as ORB predicts all chemical shifts for any given query protein. To obtain a proper match between observed peaks in the NMR experiment and predicted shifts, it is necessary to calculate from the predicted shifts a set of expected peaks for the query protein.

Expected peaks are calculated based on four inputs. The first is a set of rules describing the NMR experiment used. The PROCESS set of rules contains the most common NMR experiments, and allows addition of other experiments. The second input is a table of amino acid properties. This contains amino acid names and their bonded atoms, and gives the atoms in fast and slow exchange. Also used is a general parameter file, which allows the user to customize program variables. Finally, a table of predicted chemical shifts is used, generated from a program such as ORB.

Step 2 – calculating individual scores

Individual scores are calculated for each observed peak and expected peak pair to give a ranking from

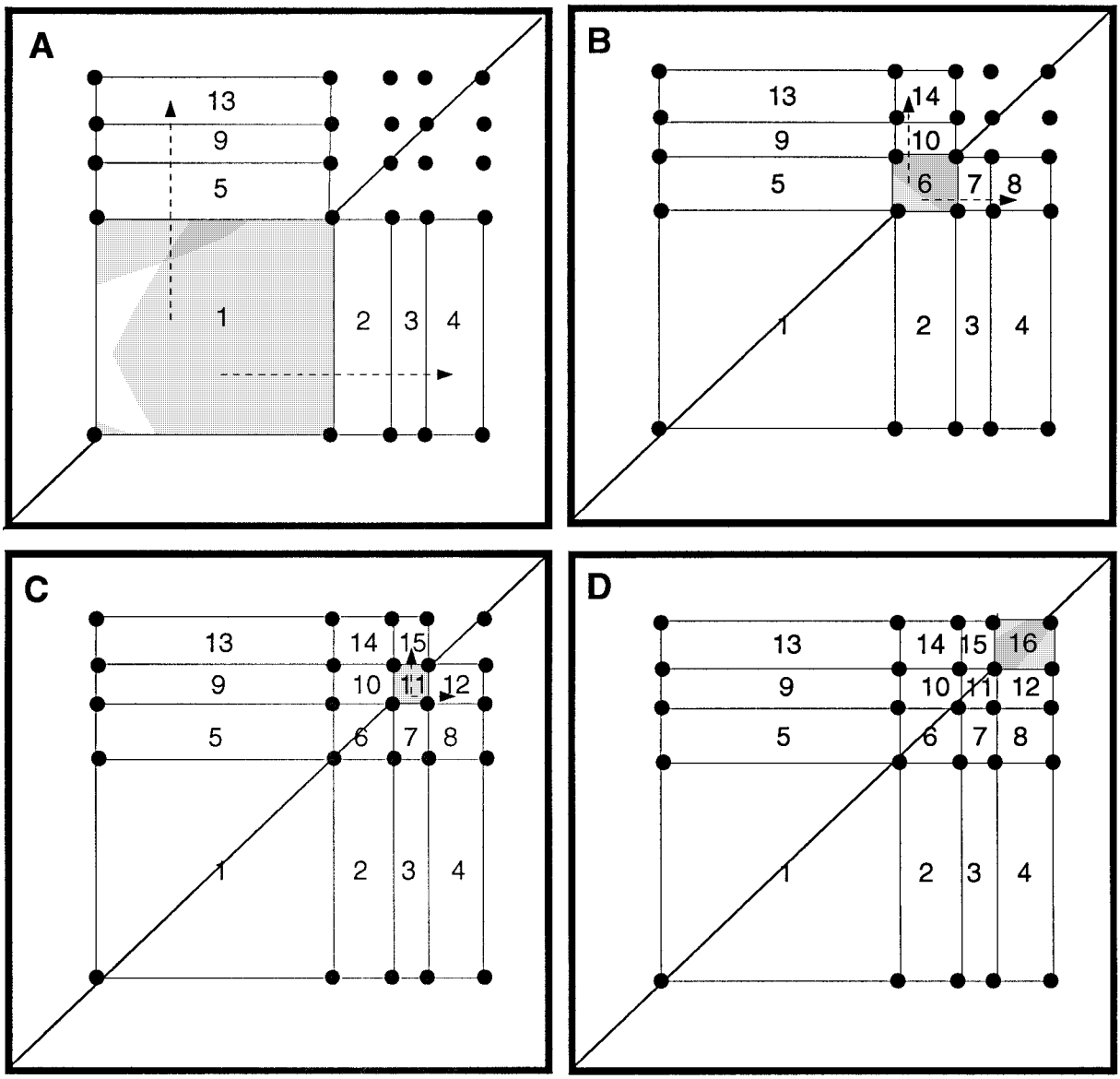

Figure 2. Steps taken in the SSS algorithm. The algorithm is explained in detail in the text.

which the best match may be determined. Each possible observed-expected pair is given a score using the following formula:

$$\exp - \left(\left(\frac{\text{observed } (x) \; - \; \text{predicted } (x)}{\text{sdev(predicted } (x))} \right)^2 + \left(\frac{\text{observed } (y) - \text{predicted } (y)}{\text{sdev(predicted } (y))} \right)^2 + \left(\frac{\text{observed}(z) - \text{predicted}(z)}{\text{sdev(predicted}(z))} \right)^2 \right) \qquad (1)$$

Observed(*x/y/z*) and predicted(*x/y/z*) are the chemical shift values (in ppm) of a particular predicted or observed peak in the *x*, *y* and *z* dimensions. The standard deviations for the predicted chemical shifts in each dimension are calculated by ORB (Gronwald et al., 1997) and serve as a measure for the expected precision of the predictions.

Step 3 – Scoring spin systems and residues

This part of PROCESS classifies spin systems into residue type and sequence location. This algorithm

takes each residue in the input sequence and compares it against each group of observed peaks (spin systems). This is done in three separate parts. The first part takes the set of predicted peaks for a residue and tries to find the best match to the set of peaks in the observed spin system. This algorithm is heuristic – not exhaustive – and will usually make the best match in a short period of time. In the next pass, the predicted peaks in slow exchange are considered. For example for a CH_2 group two expected peaks $H\beta^1$ and $H\beta^2$ have been calculated and both are in slow exchange. If one of these peaks did not match to any spin system member, then it is assigned to the same observed peak as its partner peak. The justification for this is that sometimes only one peak is observed for two expected peaks in slow exchange. In the final step, a total score for each residue/spin system pair is computed based on (1) the average score of each residue/spin system pair, (2) the percentage of spin system members which were used, and (3) the percentage of expected peaks which matched. For the case of arginine and lysine, two spin systems are scored, one originating from the amide and the other from the sidechain. A detailed explanation of the scoring may be found in the help pages of the program manual.

Step 4 – Final Output

It is possible to display the output from PROCESS in many different formats and any screen output may be saved to a file. In particular, a user may request to:

- Show Expected Peaks
- List Spin Systems
- Show Summary of Results for Residues
- Show Summary of Results for Spin Systems
- Look at Individual Peaks

PROCESS offers several different choices for any given residue. These choices are ranked by their total scores. In addition, a confidence value is calculated for all choices. Together with the total scores, the confidence values help the user to judge if an assignment is correct. For example, a low total score and a high confidence value would indicate that the match between expected and observed peaks is not very good, but that this is the only choice that fits the experimental data. If, on the other hand, the total score and the confidence value are both high, then it is almost certain that the assignment is correct.

Heteronuclear and three-dimensional capabilities

The CAMRA suite of programs are versatile and have been written to analyze peaks from heteronuclear and 3D experiments like HSQC, DQF-COSY, 3DTOCSY-HSQC spectra. The predictions provided by ORB include a complete set of ^1H, ^{13}C and ^{15}N shifts. Separation of the spin systems in a 3D experiment is carried out by the program SSS-3D, which can be accessed via the CAMRA GUI. Due to the limited amount of overlap present in, for example, ^1H-^{15}N 3D TOCSY-HSQC spectra, we found it was possible to use a simple algorithm, similar to the one used in the second stage of the SSS program. SSS-3D combines all peaks that share, within specified discrimination values, the same H^N and N frequency into one spin system. SSS-3D accepts input files in both VNMR and .PCK format (Garrett et al., 1991). CAMRA provides file conversion tools to convert other peak list formats into the .PCK format. PROCESS can calculate expected peaks for any kind of multidimensional NMR experiment and can match these with the corresponding observed spin systems.

Results

Two proteins of the chemokine family were investigated to evaluate to what extent CAMRA was able to obtain correct residue-specific assignments independent of NOE information: an IL-8 analog with a single point mutation of cysteine-7 to homocysteine, and SDF-1. To test the performance we compared the proposed residue specific assignments from PROCESS with manually derived ones for both the IL-8 analog and SDF-1. Manual assignments for both proteins were obtained using standard procedures (Wüthrich, 1986). The overall results are given for each protein (from PROCESS output), as well as comments on the performance of ORB and CAPTURE.

IL-8 analog

The ORB chemical shift predictions for this protein were based on a database of 8 highly homologous previously assigned proteins (Table 1). The results indicate that the predictions are very close to the observed shifts with average errors of 0.08 and 0.03 ppm for the N^H and H^α shifts respectively. Only in the regions around the mutation were small deviations between observed and predicted chemical shifts observed (Figure 3A). Overall the quality of the chemical

shift predictions was very good. This is mainly due to the high level of sequence identity between the IL-8 analog and members of the database of previously assigned homologues.

The CAPTURE input was generated using a 2D ^{1}H TOCSY NMR spectrum collected in H_2O at 40 °C. Water suppression and spin-lock were achieved by use of the WATERGATE (Piotto et al., 1992) and DIPSI (Shaka et al., 1988) pulse sequences, respectively. A peak list containing 1143 peaks was generated. The SSS program was used to separate the peak list into the various spin systems yielding a total of 91 spin systems. The difference between the number of spin systems (91) and the number of residues (69) is explained by the following factors. Separate spin systems are generated for the Arg and Lys main and side-chains, and several times the same spin system was broken up into two smaller incomplete spin systems due to partial overlap in the spectrum and missing peaks. For the same reasons some of the computer generated spin systems contain extra peaks which belonged to another spin system and some of them are incomplete. The IPP program allows rapid manual inspection of all spin systems to correct for these imperfections. However, to fully assess the performance of the whole CAMRA package, no corrections were performed for any of the spin systems.

A total of 63 of 69 residues were used to test the correctness of PROCESS. Two of the residues were not assigned or were invisible, and the other four residues not used were prolines. CAPTURE was used in a way that spin systems were only generated for residues which possess two or more off diagonal peaks in the fingerprint region and therefore no spin systems were generated for proline residues. As described in the Methods Section, PROCESS offers several ranked assignment choices. In testing, only the top three choices were considered. PROCESS results for the IL-8 analog are displayed in Figure 4A. Using only the predicted shifts and just one 2D TOCSY spectrum, PROCESS was able to correctly assign 67% (42/63) of all residues using only the first choice. An additional 17% (11/63) were correctly assigned using the second and third choices. In total, PROCESS correctly assigned 84% (53/63) of all residues. Overall, PROCESS was able to suggest accurate residue-specific assignment for the IL-8 analog. A detailed analysis of PROCESS results showed that there are two main causes for incorrect assignments: first, in the region of the mutation the predicted chemical shifts are not close to the observed shifts. Second, for some residues the spin systems were broken up by CAPTURE into incomplete smaller spin systems. This was particularly true for residues where peaks were obscured by the water signal. For these spin systems it was almost impossible for PROCESS to find the correct assignments.

SDF-1

To generate predicted shifts, ORB used a database of 7 previously assigned chemokines for SDF-1. In contrast to the IL-8 analog, this database does not contain any protein which possesses more than 30% sequence identity to SDF-1 (Table 1). Previous tests of ORB have shown that ORB requires at least 30% sequence identity between the query protein and one or more previously assigned proteins to produce reasonable results. Trying to assign SDF-1 was therefore a challenging test for the CAMRA package of programs. The results of the ORB predictions for SDF-1 (Figure 3B) show average errors of 0.31 and 0.23 ppm for the H^N and H^α shifts, respectively. Large errors of up to 1.5 ppm are observed for both the H^N and H^α shifts in the region between Ser-16 to Ala-21. In other regions of the protein the predictions are reasonably good for both the H^N and H^α shifts.

For CAPTURE, ^{1}H TOCSY spectra measured in H_2O as well as in D_2O were available. All spectra were measured under exactly the same conditions at 40 °C. The same peak picking strategy as described for the IL-8 analog was used for SDF-1. The only difference was the fingerprint region, where the H_2O spectrum was used, while for the rest of the spectrum the D_2O spectrum was used. By combining the H_2O and D_2O peak lists into one, it was possible to obtain a peaklist where no peaks were obscured by the water signal. The combined peak list contained a total of 1004 peaks. The SSS program separated this peak list into 61 spin systems. A comparison between the two test systems shows that the CAPTURE results for SDF-1 are better than the results for the IL-8 analog. This difference can be explained by the fact that the combination of D_2O and H_2O TOCSY spectra allowed for the creation of a SDF-1 peak list where no signals were obscured by the water signal. The high quality of the SDF-1 peak list enabled CAPTURE to create an almost perfect set of spin systems. It is important to note that the quality of the automatically generated spin systems strongly depends on the amount of overlap which is present in the region of interest of the spectrum.

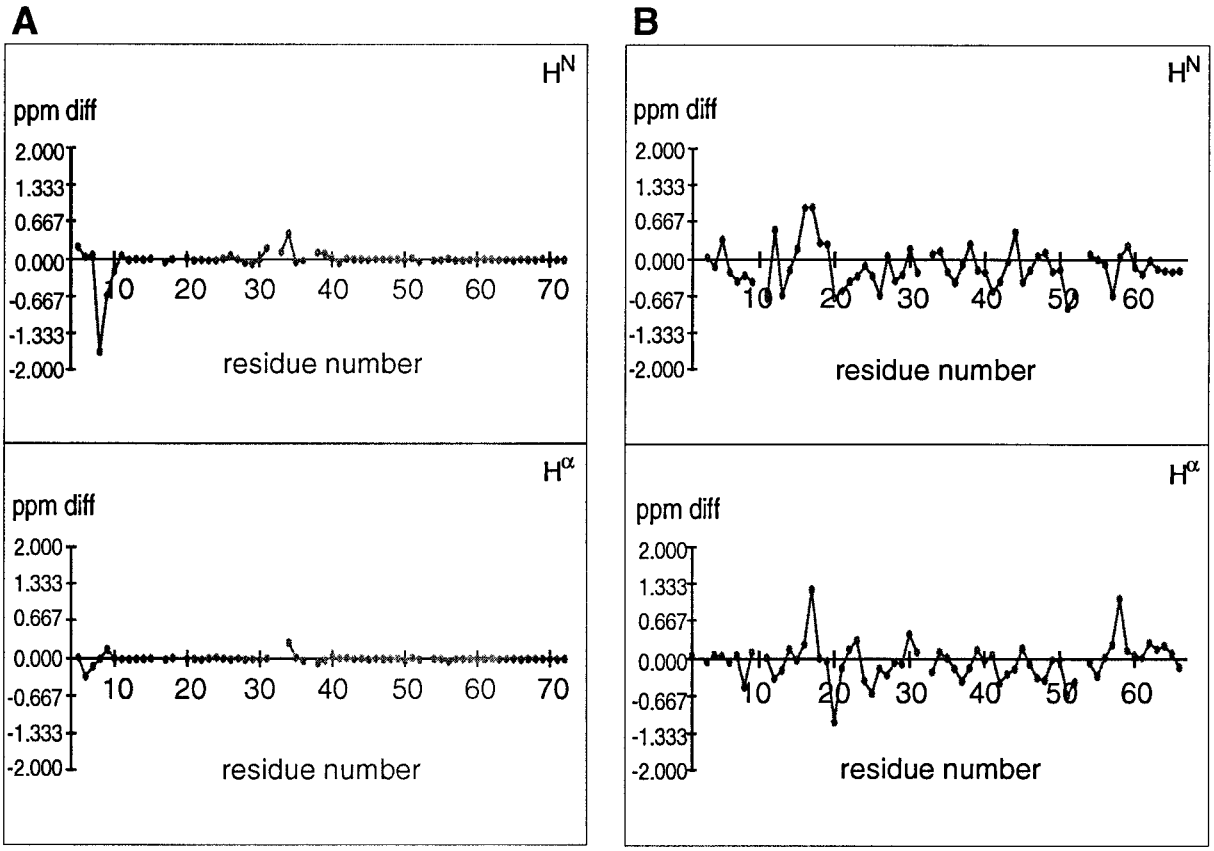

Figure 3. Results from the chemical shift predictions of the backbone H^N and H^α nuclei of the IL-8 analog (A) and SDF-1 (B) plotted as a function of residue number. The diagrams show observed shifts minus predicted shifts. The upper panel contains the results for H^N shifts whereas the lower panel contains the corresponding results for the H^α shifts. The diagrams were created using the graphical shift comparison program GSC (Gronwald et al., 1997).

Manual inspection of the spectra and a comparison with the available residue-specific assignments showed that for several residues no peaks could be found and in contrast to the IL-8 analog only for very few residues spin systems were broken up into smaller incomplete spin systems. Consequently the number of spin systems generated by CAPTURE was smaller than the number of residues. It is important to note, that for both the IL-8 analog and SDF-1, the CAPTURE parameters were set such that two or more off-diagonal peaks in the fingerprint region were required for each spin system.

For SDF-1, a total of 56 out of 67 residues were used to test the correctness of PROCESS. Seven of the residues were invisible in the TOCSY spectra and 4 were prolines for which no CAPTURE spin systems were generated. Consequently, the performance of PROCESS was evaluated on the basis of the available 56 residues. The results for SDF-1 are displayed in Figure 4B. Using the top three choices PROCESS was able to correctly assign 52% (29/56) of all residues.

Manual intervention

To fully assess the performance of PROCESS a set of hand edited 'ideal' spin systems was provided and the SDF testing was repeated. These spin systems were generated using the fingerprint region of the H_2O TOCSY spectrum and resolving all chemical shift degeneracy and overlap manually by an expert user. Using this set of 'ideal' spin systems PROCESS was able to correctly assign 57% (32/56) of all residues with the first three choices. If the SDF-1 shifts themselves were used as predicted shifts, PROCESS was able to correctly assign 100% (56/56) of the residues.

Table 1. Database of previously assigned homologous sequences of the chemokine family.

Sequence	% identity to IL-8(4-72)C7HC	% identity to SDF (1–67)	Reference
IL-8 (1-72)	100		Rajarathnam et al., 1994
IL-8 (4-72)	100		Rajarathnam et al., 1994
IL-8 (5-72)	99		Rajarathnam et al., 1994
IL-8 (6-72)	97		Rajarathnam et al., 1994
IL-8 H33A (4-72)	99		Rajarathnam et al., 1994
IL-8 E38A (4-72)	99		Rajarathnam et al., 1996
IL-8 I10A (4-72)	99		Rajarathnam et al., 1994
IL-8 R6K (4-72)	99		Rajarathnam et al., 1994
IL-8 L25N (4-72)		30	Rajarathnam et al., 1995
MGSA (1-72)		28	Kim et al., 1994
PF4-M2 (1-67)		27	Mayo et al., 1995
RANTES		28	Skelton et al., 1995
MCP-3		21	Kim et al., 1996
MIP-1β		25	Lodi et al., 1994
MCP-1		18	Handel et al., 1996

Discussion

Recent advances in molecular biology and NMR methodology have resulted in an explosion in the number of protein structures solved by NMR spectroscopy. Complete or nearly complete chemical shift assignments for more than 200 proteins have been deposited in the BioMagResBank (Seavey et al., 1991), and this number is likely to increase with passing time. This chemical shift database can provide a basis for automated assignment of mutant proteins and other related proteins using programs like CAMRA. Hereby, the sequence homology between query and previously assigned proteins can vary largely (∼30% to 99%). The CAMRA approach is unique in that it does not make use of any NOESY data. Instead, CAMRA assigns chemical shifts in a residue specific manner based on predicted chemical shifts and TOCSY spectra. The program is versatile and time effective providing the chemical shifts of at least one related protein are available in the database and the cross peaks in the TOCSY spectrum are well resolved.

In order to test the utilities and limitations of this approach, CAMRA was used to obtain residue-specific assignments of two query proteins. The first protein tested was an interleukin-8 analog having a single mutation compared to the native protein. For this analog, chemical shift assignments of a number of highly homologous (≥95%) IL-8 mutants were available in the database. The second protein tested was SDF-1 and in this case, the sequence similarity was in the order of 18–30%. It was observed that the chemical shift predictions were substantially better for the IL-8 analog than for SDF-1. Nevertheless, 52% of the residues could be assigned using the computer generated spin systems. This information should aid in assigning the rest of the protein from other NMR experiments.

These results show clearly that reasonable results can be obtained using CAMRA even in the case where a database of relatively low homologous proteins is available for the ORB prediction process. A comparison of the following SDF-1 results gives an indication which factors are important in obtaining good PROCESS results. Using the computer generated spin systems and the regular ORB predictions 52% of the residues could be assigned. If hand edited 'perfect' spin systems from the fingerprint region were used, this percentage increased to 57%. In the next test the SDF-1 shifts themselves were used together with the computer generated spin systems and 75% of all residues could be assigned. 100% of all residues could be assigned with the first three choices using the 'perfect' spin systems and the SDF-1 shifts themselves. Together with the IL-8 results it becomes clear that the most important factor in obtaining a correct assignment is an accurate set of predicted shifts, closely followed by the quality of the generated spin systems.

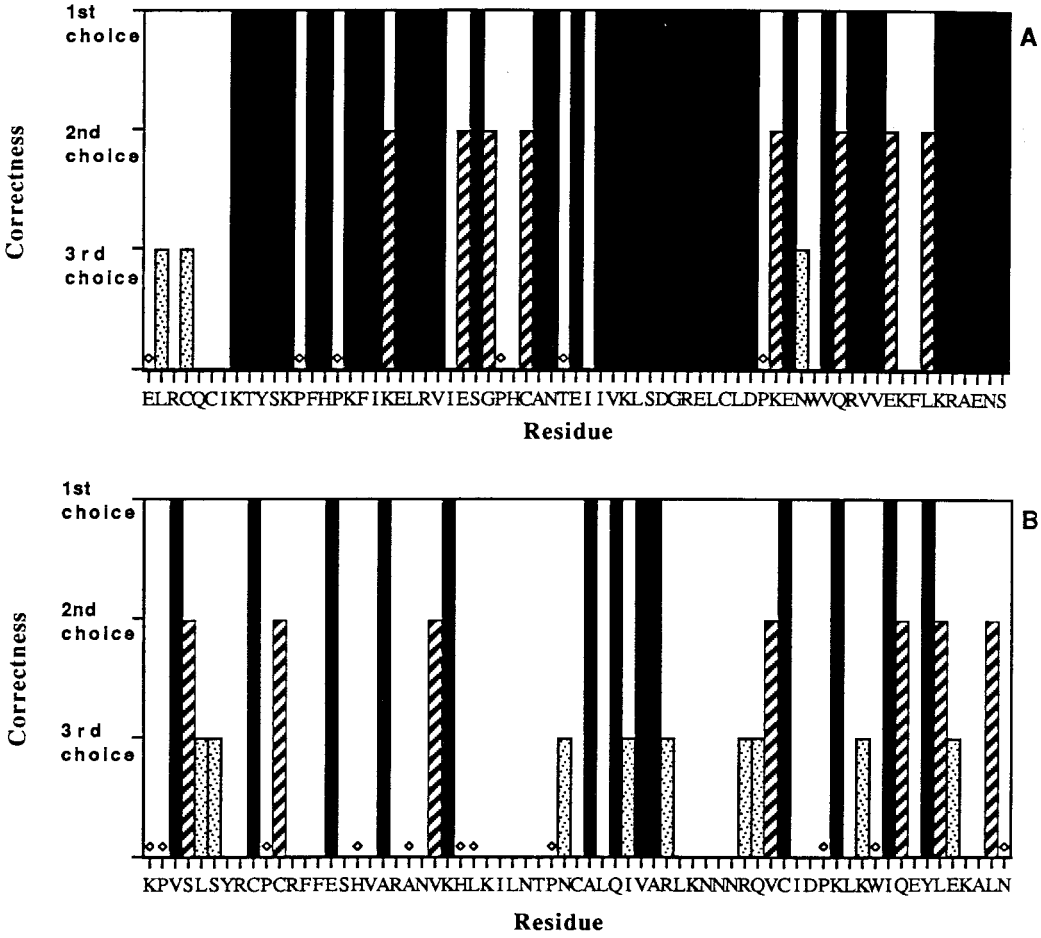

Figure 4. PROCESS results for the SDF-1 (A) and for the IL-8 analog (B). The columns display the level of correctness. PROCESS gives its top three ranked choices for each assignment. A solid bar shows when the first choice, a striped bar shows when the second choice, and a shaded bar shows when the third choice is correct. If no bar is given, PROCESS was not able to correctly make an assignment with its top three choices. A diamond shows when no information is available for a residue, so PROCESS cannot be rated on these residues.

Even if a high percentage of the assignments is correct it is important for the user to find out which of the suggested assignments are most likely correct and which are probably wrong. It was found that the combination of score values and confidence levels which are both calculated by PROCESS is really helpful in this regard.

We would like to point out that the assignment success is a function of the quality of the TOCSY spectrum, chemical shift dispersion and the extent of the homology of the proteins in the data base. Poor water suppression with loss of peaks near the H_2O region will substantially effect the quality of the TOCSY spectrum and accordingly the output from CAPTURE. We obtained the best results by the use of gradients for H_2O suppression (WATERGATE pulse sequences). CAMRA is particularly useful when a series of proteins with single or multiple mutations are being studied. The assignments of each protein is added to the database which further aids the assignments of the subsequent proteins. CAMRA is also useful in assigning flexible regions of the protein which are hard to assign using conventional methods due to lack of NOEs. Due to practical reasons CAMRA was tested on medium sized proteins. However, CAMRA should be applicable to larger proteins as well providing that good quality TOCSY spectra are available. The three CAMRA parts (PROCESS, ORB, and CAPTURE) can be used combined together in one package, or used separately to help the user in the various stages of a manual assignment.

In summary, CAMRA is a very versatile program package which should be useful for a wide variety of applications relating to a residue specific protein assignment.

Availability

The complete CAMRA suite of executables for a Sun or SGI may be accessed from the following WEB page: http://www.pence.ualberta.ca/ftp.

Acknowledgements

We wish to thank Matthew Crump for access to the SDF-1 chemical shift data, Colin Bigam for the testing of the programs and Bruce Lix for many helpful discussions. In addition we would like to thank all members of the Brian Sykes lab for their contributions to the paper. This work is supported by the Protein Engineering Network of Centres of Excellence of Canada and the Alberta Heritage Foundation for Medical Research.

References

Baggiolini, M., Dewald, B. and Moser, B. (1996) *Annu. Rev. Immunol.*, **15**, 675–705.

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

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

Bigam, C., Jellard, T., Gronwald, W. and Sykes, B.D. (1998) *Magnetic Moments*, **9**, 9–12.

Clore, G.M. and Gronenborn, A.M. (1991) *Science*, **252**, 1390–1399.

Croft, D., Kemmink, J., Neidig, K.-P. and Oschkinat, H. (1997) *J. Biomol. NMR*, **10**, 207–219.

Crump, M.P., Gong, J.-H., Loetscher, P., Rajarathnam, K., Arenzana-Seisdedos, F., Virelizier, J.-L., Baggioloini, M., Sykes, B. D. and Clark-Lewis, I. (1997) *EMBO J.*, **16**, 6996–7007.

de Dios, A.C., Pearson, J.G. and Oldfield, E. (1993) *Science*, **5113**, 1491–1496.

Fairbrother, W.J. and Skelton, N.J. (1996) *Three dimensional structures of the chemokine family*. In R. Horuk (ed.), Chemoattractant Ligands and their Receptors, CRC Press, London, pp. 55–86.

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

Garrett, D.S., Powers, R., Gronenborn, A. and Clore, G.M. (1991) *J. Magn. Reson.*, **94**, 214–220.

Gronwald, W., Boyko, R.F., Sönnichsen, F.D., Wishart, D.S. and Sykes, B.D. (1997) *J. Biomol. NMR*, **10**, 165–179.

Gronwald, W., Boyko, R.F. and Sykes, B.D. (1997) *CABIOS*, **13**, 557–558.

Handel, T.M. and Domaille, P.J. (1996) *Biochemistry*, **33**, 15283–15292.

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

Kim, K.-S., Rajarathnam, K., Clark-Lewis, I. and Sykes, B.D. (1996) *FEBS Lett.*, **395**, 277–282.

Kjaer, M., Andersen, K.V. and Poulsen, F.M. (1994) *Methods Enzymol.*, **239**, 288–318.

Kleywegt, G.J., Boelens, R., Cox, M., Llinás, M. and Kaptein, R. (1991) *J. Biomol. NMR*, **1**, 23–47.

Kraulis, P.J. (1994) *J. Mol. Biol.*, **243** 696–718.

Lodi, P.J., Garret, D.S., Kuszewski, J., Tsang, M.L.-w., Weatherbee, J.A., Leonard, W.J., Gronenborn, A.M. and Clore, G.M. (1994) *Science*, **263**, 1762–1767.

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

Meadows, R.P., Olejniczak, E.T. and Fesik, S.W. (1994) *J. Biomol. NMR*, **4**, 79–96.

Morelle, N., Brutscher, B., Simorre, J-P. and Morelle, M.D. (1995) *J. Biomol. NMR*, **5**, 154–160.

Olson, J.B. Jr. and Markley, J.L. (1994) *J. Biomol. NMR*, **4**, 385–410.

Oschkinat, H., Holak, T.A. and Cieslar, C. (1991) *Biopolymers*, **31**, 699–712.

Ousterhout, J.K. (1994) *Tcl and the Tk Toolkit*, Addison-Wesley Professional Computing Series.

Piotto, M., Saudek, V. and Sklenar, V. (1992) *J. Biomol. NMR*, **2**, 661–665.

Rajarathnam, K., Clark-Lewis, I. and Sykes, B.D. (1994) *Biochemistry*, **33**, 6623–6630.

Rajarathnam, K., Clark-Lewis, I. and Sykes, B.D. (1995) *Biochemistry*, **34**, 12983–12990.

Rajarathnam, K., Clark-Lewis, I., Dewald, B., Baggiolini, M. and Sykes, B.D. (1996) *FEBS Lett.*, **399**, 43–46.

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

Shaka, A.J., Lee, C.J. and Pines, A. (1988) *J. Magn. Reson.*, **77**, 274–293.

Skelton, N.J., Aspiras, F. and Schall, T.J. (1995) *Biochemistry*, **34**, 5329–5342.

Wall, L. and Schwartz, R.L. (1990) *Programming perl*, O'Reilly and Associates, Inc.

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

Xu, J. and Sanctuary, B.C. (1993) *J. Chem. Inf. Comput. Sci.*, **33**, 490–500.

Xu, J., Weber, P.L. and Borer, P.N. (1995) *J. Biomol. NMR*, **5**, 183–192.

Zimmermann, D., Kulikowski, C., Wang, L., Lyons, B. and Montelione, G.T. (1994) *J. Biomol. NMR*, **4**, 241–256.