Journal of Biomolecular NMR, **17:** 283–294, 2000.
KLUWER/ESCOM
© 2000 *Kluwer Academic Publishers. Printed in the Netherlands.*

DipoCoup: A versatile program for 3D-structure homology comparison based on residual dipolar couplings and pseudocontact shifts

Jens Meiler[a], Wolfgang Peti[a] & Christian Griesinger[a,b,*]
[a]*Universität Frankfurt, Institut für Organische Chemie, Marie-Curie-Str. 11, D-60439 Frankfurt am Main, Germany*
[b]*Max-Planck Institute for Biophysical Chemistry, Am Faßberg 11, D-37077 Göttingen, Germany*

Received 22 Februuary 2000; Accepted 28 April 2000

Key words: bioinformatics, dipolar couplings, postgenomic research, protein data base, structure elucidation, three dimensional homology search

Abstract

A program, *DipoCoup*, is presented that allows to search the protein data bank for proteins which have a three dimensional fold that is at least partially homologous to a protein under investigation. The three dimensional homology search uses secondary structure alignment based on chemical shifts and dipolar couplings or pseudocontact shifts for the three dimensional orientation of secondary structure elements. Moreover, the program offers additional tools for handling and analyzing dipolar couplings.

Introduction

One goal of post genomic research is to determine all protein folds. The number of folds is expected to be limited (Sali, 1998; Fischer and Eisenberg, 1999). Sequence profile methods nowadays have a big impact in fold recognition. *Ab initio* structure prediction works up to 40–60 amino acids and may emerge as a powerful tool for structure prediction in the future (Moult, 1999). To obtain a complete coverage of folds most effectively, it is important to focus on the elucidation of structures with novel folds rather than rediscovering known folds on new proteins. Blast threading and *ab initio* approaches rely on the analysis of primary and secondary structure in the context of a three dimensional structure database. We will present experimental tools that allow to compare the 3D fold of a new protein to all known folds in an early stage of NMR based structure determination. This approach has the potential to predict folds of a new protein with little homology to proteins with known folds. By the same token, structure elucidation of a new protein with a structure homologous to a known fold will be accelerated. There is so far only one example of using experimental NMR parameters in an early stage for 3D homology searches (Annila et al., 1999). Recently the possibility for using protein fragments generated from PDB and chosen by aligning similar dipolar couplings and chemical shifts for structure determination was shown (Delaglio et al., 2000). The availability of orientation information from NMR experiments in terms of residual dipolar couplings (Tolman et al., 1995; Tjandra and Bax, 1997; Bax and Tjandra, 1997; Clore et al., 1998a; Fischer et al., 1999; Peti and Griesinger, 2000; Meiler et al., 2000) offers new possibilities in this field. In this paper, we present a versatile program, *DipoCoup*, that uses chemical shifts for the alignment of secondary structure elements and tertiary structure alignment from dipolar couplings and pseudocontact shifts for the homology search in the PDB. We will show, using examples, that the program is fast enough to search through a large number of pdb files.

[*] To whom correspondence should be addressed. E-mail: cigr@org.chemie.uni-frankfurt.de

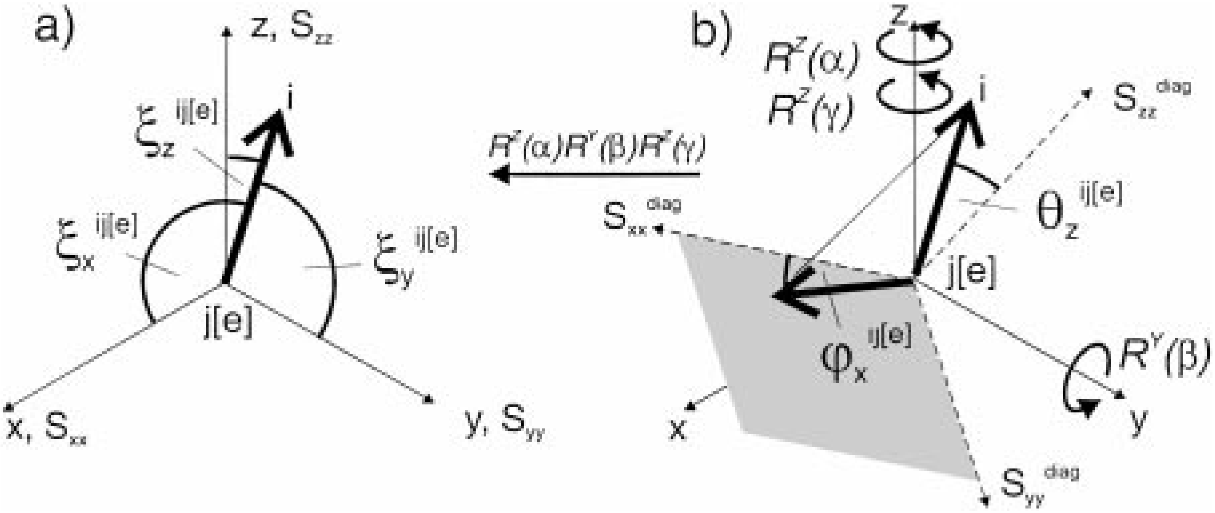

Figure 1. (a) Coordinate system of the molecule *(x, y, z)* with a bond vector between the two nuclei *i* and *j* or a nucleus *i* and an electron *e*. The projection angles of the vector onto the *x, y, z* axes are ξ_x^{ij}, ξ_y^{ij} and ξ_z^{ij}, respectively. (b) Representation of the vector in the frame of the tensor S_{xx}^{diag}, S_{yy}^{diag}, S_{zz}^{diag}. The Euler rotation transforms the tensor into the coordinate system of the molecule. The orientation of the bond vector $\vec{r}_{ij}$ is defined by the angles θ_z^{ij} and φ_x^{ij}.

Theory

Experimental dipolar couplings between nuclei *i* and $j(D^{ij})$ and pseudocontact shifts between nucleus *i* and electron $e(\delta_{PC}^{ie})$ are related to the alignment tensor (principal components: A_{xx}, A_{yy}, A_{zz}) or to the magnetic susceptibility tensor (principal components: χ_{xx}, χ_{yy}, χ_{zz}) and to the orientation of a specific vector with respect to the alignment tensor expressed by the projection angles θ_z^{ij} and φ_x^{ij} according to Equation 1. The vector is either the vector $\vec{r}_{ij}$ between the two coupled atoms *i* and *j* in case of dipolar coupling (Equation 1a) or the vector $\vec{r}_{ie}$ between the electron spin *e* (paramagnetic center) and the active nucleus *i* (Equation 1b).

$$D^{ij}(\theta_z^{ij}, \varphi_x^{ij}) = \frac{-\mu_0 h S \gamma_i \gamma_j}{8\pi^3 r_{ij}^3}\left[\frac{1}{6}(2A_{zz} - A_{xx} - A_{yy})(3\cos^2\theta_z^{ij} - 1) + \frac{1}{2}(A_{xx} - A_{yy})\cos 2\varphi_x^{ij}\sin^2\theta_z^{ij}\right], \tag{1a}$$

$$\delta_{PC}^{ie}(\theta_z^{ie}, \varphi_x^{ie}) = \frac{10^6}{6\pi r_{ie}^3}\left[\frac{1}{6}(2\chi_{zz} - \chi_{xx} - \chi_{yy})(3\cos^2\theta_z^{ie} - 1) + \frac{1}{2}(\chi_{xx} - \chi_{yy})\cos 2\varphi_x^{ie}\sin^2\theta_z^{ie}\right] \tag{1b}$$

In case of paramagnetic alignment the susceptibility tensor $\hat{\chi}$ is related to the alignment tensor $\hat{A}$ by $\hat{\chi} = \hat{A}(15\mu_0 kT/4B_0\pi)$. Equation 1 uses the alignment tensor as the frame of reference (Figure 1). The formal dependence of dipolar couplings and pseudo contact shifts is the same while the prefactors differ. The prefactor is constant for dipolar couplings (Equation 1a) if the distance r_{ij} is constant. For pseudocontact shifts and also for dipolar couplings between nuclei whose distance is not fixed a priori in the bonding network they vary because the distance r_{ie} or r_{ij} cannot be regarded as constant in this case (Ghose and Prestegard, 1997; Clore and Garrett, 1999).

However, the measured orientation value cannot be translated directly in a combination of θ_z^{ij} and φ_x^{ij}. An infinite number of combinations of θ_z^{ij} and φ_x^{ij} exist, that fulfill an experimental value. Still if one pair of angles θ_z^{ij} and φ_x^{ij} can be found to be correct due to the alignment of a whole molecule, four orientations of the molecule fulfill all experimental values, since the signs of angles θ_z^{ij} and φ_x^{ij} can be reversed independently in Equation 1 without the change of either dipolar couplings or pseudo contact shifts.

In the context of a 3D homology search, the coordinate system of a protein in the 3D structure data file (e.g. PDB) is the natural frame of reference. Therefore we express Equation 1 in this coordinate system which is rotated by three Euler angles α, β, γ with respect to the alignment tensor (Figure 1). Equation 2 expresses the dipolar couplings in the molecular frame (an identical equation is obtained for pseudocontact shifts δ_{PC}^{ie} by replacing j with e, all following equations are only given for dipolar couplings):

$$
\begin{aligned}
D^{ij}(\xi_x^{ij}, \xi_y^{ij}, \xi_z^{ij}) &= \\
&= F_{ij}\begin{pmatrix} \cos\xi_x^{ij} \\ \cos\xi_y^{ij} \\ \cos\xi_z^{ij} \end{pmatrix}^T \begin{pmatrix} -S_{yy}-S_{zz} & S_{xy} & S_{xz} \\ S_{xy} & S_{yy} & S_{yz} \\ S_{xz} & S_{yz} & S_{zz} \end{pmatrix} \begin{pmatrix} \cos\xi_x^{ij} \\ \cos\xi_y^{ij} \\ \cos\xi_z^{ij} \end{pmatrix} \\
&= F_{ij}\begin{pmatrix} (\cos^2\xi_y^{ij} - \cos^2\xi_x^{ij})S_{yy} + (\cos^2\xi_z^{ij} - \cos^2\xi_x^{ij})S_{zz} \\ +(2\cos\xi_x^{ij}\cos\xi_y^{ij})S_{xy} + (2\cos\xi_x^{ij}\cos\xi_z^{ij})S_{xz} \\ +(2\cos\xi_y^{ij}\cos\xi_z^{ij})S_{yz} \end{pmatrix}
\end{aligned}
\tag{2}
$$

with

$$F_{ij} = \frac{-\mu_0 h S \gamma_i \gamma_j}{8\pi^3 r_{ij}^3}$$

In this molecular frame the alignment tensor is no longer diagonal and can be expressed by a symmetric three by three traceless matrix holding five independent elements S_{xx}, S_{zz}, S_{xy}, S_{xz} and S_{yz}, the elements of the Saupe matrix (Saupe, 1968). The eigenvalues of this matrix S_{xx}^{diag}, S_{yy}^{diag}, S_{zz}^{diag} are identical to the principal components of the alignment tensor A_{xx}, A_{yy}, A_{zz}. The angles ξ_x^{ij}, ξ_y^{ij}, ξ_z^{ij} define the projection angles of the bond vector $\vec{r}_{ij}$ or the vector between the nucleus and the electron $\vec{r}_{ie}$ using pseudocontact shifts onto the molecular frame. For a given structure and experimental dipolar couplings D_{exp}^{ij}, the five independent tensor contributions can be determined directly by solving the linear system of equations given from Equation 3 for a set of experimental dipolar couplings for n pairs of nuclei i and j requiring $D_{\text{exp}}^{ij} = D_{\text{theor}}^{ij}$ (Losonczi et al., 1999).

$$
\begin{pmatrix} D_{\text{exp}}^{ij1}/F_{ij} \\ \vdots \\ D_{\text{exp}}^{ijn}/F_{ij} \end{pmatrix} \overset{!}{=} \begin{pmatrix} D_{\text{theor}}^{ij1}/F_{ij} \\ \vdots \\ D_{\text{theor}}^{ijn}/F_{ij} \end{pmatrix} =
$$

$$
\begin{pmatrix}
\cos^2\xi_y^{ij1} - \cos^2\xi_x^{ij1} & \cos^2\xi_y^{ij1} - \cos^2\xi_x^{ij1} & 2\cos\xi_x^{ij1}\cos\xi_y^{ij1} & 2\cos\xi_x^{ij1}\cos\xi_y^{ij1} & 2\cos\xi_x^{ij1}\cos\xi_y^{ij1} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\cos^2\xi_y^{ijn} - \cos^2\xi_x^{ijn} & \cos^2\xi_y^{ijn} - \cos^2\xi_x^{ijn} & 2\cos\xi_x^{ijn}\cos\xi_y^{ijn} & 2\cos\xi_x^{ijn}\cos\xi_y^{ijn} & 2\cos\xi_x^{ijn}\cos\xi_y^{ijn}
\end{pmatrix}
\begin{pmatrix} S_{yy} \\ S_{zz} \\ S_{xy} \\ S_{xz} \\ S_{yz} \end{pmatrix} = \mathbf{C}\vec{S}
\tag{3}
$$

This system of equations can be solved by multiplication of the pseudo inverse of the rectangular matrix **C**, i.e., by calculating the Moore-Penrose-Inverse of the matrix yielding the vector $\vec{S}$. Rebuilding the Saupe matrix from these values and analyzing its eigensystem yields the eigenvalues of the tensor S_{xx}^{diag}, S_{yy}^{diag}, S_{zz}^{diag} as well as its orientation given by the eigenvectors. It can be expressed in terms of three Euler angles in α, β, and γ.

$$
\begin{aligned}
\begin{pmatrix} -S_{yy}-S_{zz} & S_{xy} & S_{xz} \\ S_{xy} & S_{yy} & S_{yz} \\ S_{xz} & S_{yz} & S_{zz} \end{pmatrix} &= \left(R^Z(\alpha)R^Y(\beta)R^Z(\gamma)\right)^T \\
&\quad \begin{pmatrix} S_{xx}^{\text{diag}} & 0 & 0 \\ 0 & S_{yy}^{\text{diag}} & 0 \\ 0 & 0 & S_{zz}^{\text{diag}} \end{pmatrix} \\
&\quad R^Z(\alpha)R^Y(\beta)R^Z(\gamma)
\end{aligned}
\tag{4}
$$

The solution of the Moore Penrose inversion problem is equivalent to finding a solution D_{theor}^{ij} with the least square deviation for a given experimental set of D_{exp}^{ij}. Experimental errors cannot be directly taken into consideration during this approach. Therefore a careful analysis afterwards is necessary according to Losonczi et al. (1999).

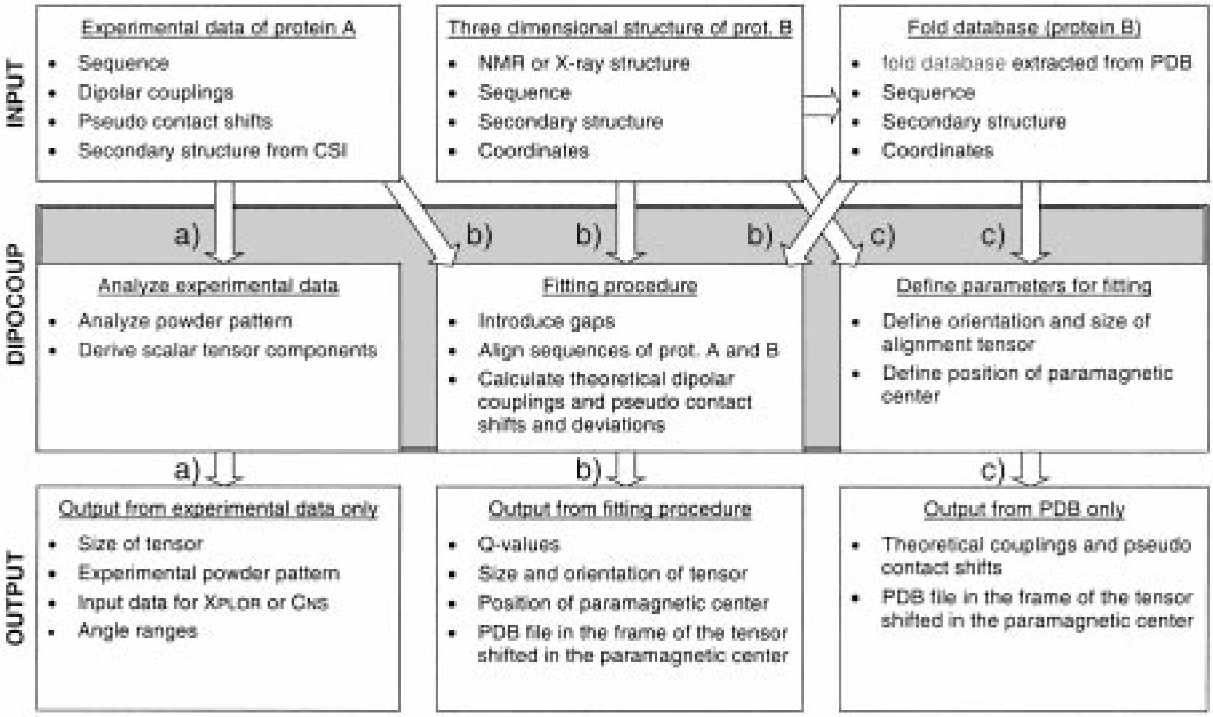
INPUT
Experimental data of protein A
• Sequence
• Dipolar couplings
• Pseudo contact shifts
• Secondary structure from CSI
Three dimensional structure of prot. B
• NMR or X-ray structure
• Sequence
• Secondary structure
• Coordinates
Fold database (protein B)
• fold database extracted from PDB
• Sequence
• Secondary structure
• Coordinates
DIPOCOUP
a)
b)
b)
b)
c)
c)
Analyze experimental data
• Analyze powder pattern
• Derive scalar tensor components
Fitting procedure
• Introduce gaps
• Align sequences of prot. A and B
• Calculate theoretical dipolar couplings and pseudo contact shifts and deviations
Define parameters for fitting
• Define orientation and size of alignment tensor
• Define position of paramagnetic center
a)
b)
c)
OUTPUT
Output from experimental data only
• Size of tensor
• Experimental powder pattern
• Input data for XPLOR or CNS
• Angle ranges
Output from fitting procedure
• Q-values
• Size and orientation of tensor
• Position of paramagnetic center
• PDB file in the frame of the tensor shifted in the paramagnetic center
Output from PDB only
• Theoretical couplings and pseudo contact shifts
• PDB file in the frame of the tensor shifted in the paramagnetic center

Figure 2. Schematic features of the program *DipoCoup*. Arrows (a) describe the analysis of experimental dipolar couplings and/or pseudocontact shifts from protein **A** without the knowledge of its three dimensional structure and without the use of the database. Arrows (c) describe the analysis of the three dimensional structure of protein **B** for calculating theoretical dipolar couplings or pseudocontact shifts. Arrows (b) indicate the fitting procedure of protein **A** to the known three dimensional structure of protein **B**. This is used to obtain the orientation of the alignment tensor derived from the experimental data for **A** in the molecular frame of protein **B**. The quality of the fit is measured by the *Q*-value. Alternatively to a single molecule **B** a whole database can be searched finding homologous structures or structure fragments.

Materials and methods

The 3D homology search program *DipoCoup* was written in C^{++} and can be run on every standard PC working with either Windows95/98 or WindowsNT. The program offers three general means of analyzing dipolar couplings and pseudocontact shifts (Figure 2) of the protein **A** under investigation by comparing it to one or several selected proteins **B** from the database. In procedure (a) one can analyze the experimental dipolar couplings and/or pseudocontact shifts of protein **A** as well as obtain secondary structure information from ^{13}C chemical shift index data, CSI (Spera and Bax, 1991; Wishart et al., 1992, 1995; Wishart and Sykes, 1994). The program is able to handle different sets of dipolar couplings in combination with pseudocontact shifts for one alignment tensor. Dipolar couplings for atom pairs with defined distances (e.g., N-H^{N}, Cα-Hα, Cα-CO) in protein **A** scaled with F_{ij}^{-1} can be visualized in a histogram yielding a powder pattern. The eigenvalues S_{xx}^{diag}, S_{yy}^{diag}, S_{zz}^{diag} of the tensor can then be determined from the histogram (Clore et al., 1998b). With this information the program can generate input files for XPLOR- or CNS-annealing protocols which use residual dipolar couplings as restraints. It also calculates the angle projection ranges that allow to use dipolar couplings in XPLOR- or CNS-calculations without the necessity to define the orientation of the alignment tensor (Meiler et al., 2000).

3D homology searching is prepared in (c) by calculating NMR properties of a potentially homologous protein **B** which is extracted from a structure data base. From the given three dimensional structure of protein **B** a set of dipolar couplings and pseudocontact shifts can be generated. To do this, the program first adds hydrogen atoms that may be missing in the structure and corrects the bond lengths between all heavy atoms and their bound hydrogen atoms according to Bax and Ottiger (1998). For a given alignment tensor and paramagnetic center theoretical dipolar couplings and pseudocontact shifts can be calculated, visualized and exported to disk, respectively. Also the three dimensional structure can be exported to disk oriented in the frame of reference of the alignment tensor and shifted to the appropriate position with respect to the paramagnetic center.

Finally in procedure (b), both the experimental data from the protein **A** under investigation and the three dimensional structure from the protein **B** can be checked for matching 3D folds. One or several proteins **B** from the protein data bank (PDB) can be used, allowing one to compose the experimental data to a database of proteins. Hydrogen atoms are added or corrected for proteins **B** as already described in (c). The secondary structure elements of proteins **B** are calculated from the coordinates by analyzing hydrogen bonds and φ- and ψ-angles. Then the alignment of residues *a* of **A** and *b* of **B** is done such that D_{exp}(a) is assigned to the respective atoms of residue *b* of protein **B**. This set of 'experimental' dipolar couplings is used to calculate the alignment tensor and its orientation according to Equation 3. In this case no analysis of the histogram needs to be performed. As a quality measure the *Q*-value of the dipolar couplings (analogous for pseudocontact shifts) is used: $Q = \sqrt{\sum_{ij}(D^{ij}_{\mathrm{exp}} - D^{ij}_{\mathrm{theor}})^2 / \sum_{ij}(D^{ij}_{\mathrm{exp}})^2}$ (Cornilescu et al., 1998). *Q* is a normalized square deviation and is equivalent to $\sqrt{2}$ times the *R*-factor (Clore and Garrett, 1999). Moreover, the program calculates the correlation coefficient R (not to be mixed up with the *R*-factor) and offers therefore a second quality value.

The alignment of the residues *a* of protein **A** and *b* of protein **B** is not based on primary sequence homology. Rather, the sequences will be aligned to have a minimum *Q*-value. The program aligns first all amino acids of protein **A** over the amino acids of protein **B** starting with the first for both proteins, respectively. After calculation of the Q factor for this alignment the sequence of protein **A** is shifted by one residue and the procedure is repeated until the last amino acid of **A** is aligned with the last amino acid of **B**. This ensures that terminal secondary structure elements of protein **A** are fully used in the alignment process. This procedure avoids to find false positive hits due to a changing number of dipolar couplings and CSI data used. A check for matches of the secondary structure elements is performed. Secondary structure elements are derived from CSI for protein **A** and by analysis of H-bonds and ϕ and ψ (procedure (c) in Figure 2) for protein **B**. To achieve optimal alignment, the secondary structure elements of **A** can be disconnected and aligned individually with matching secondary structure elements of **B**. By default, disconnection of secondary structure elements in **A** occurs at boundaries of secondary structure elements, e.g. from β-sheet to random coil. However, the user may also suggest other positions for disconnecting the sequence, if additional information has to be used or other ideas have to be tesed. If no secondary structure alignment is possible the alignment with minimal *Q* without the use of secondary structural information is presented. The program allows for a search over the whole or part of the PDB database, as will be described subsequently.

If pseudocontact shifts are given, the position of the paramagnetic center either can be explicitly defined in the three dimensional structure or can be optimized by an interactive grid search protocol. For optimization to proceed, a starting position, a starting step, and the size of the cube to be searched has to be supplied. The program searches this given cube using the starting step size and restarts this search with the best point of the previous search and a decreased step size and size of the cube, until the step size is smaller than a predefined target value (e.g., 0.1 Å).

The program can be downloaded together with an example and two databases of 125 (Rost and Sander, 1994) and 500 representative folds out of the PDB from: http://krypton.org.chemie.uni-frankfurt.de/∼mj/software.html

$^1J_{\mathrm{NH}}$ and $^1D_{\mathrm{NH}}$ couplings were measured for the protein *Hgi*CIC (C46 → S) using the direct measurement of the $^1J_{\mathrm{NH}}$ splitting in the ^{15}N dimension of 2D ^{1}H-^{15}N HSQC spectra and $^1J_{\mathrm{NH}}$ modulated spectra (Tjandra et al., 1996). To measure the dipolar couplings, two ^{15}N labeled samples of *Hgi*CIC (C46 → S) were prepared: One for measuring isotropic $^1J_{\mathrm{NH}}$ couplings and one sample where the weak alignment to the magnetic field is induced using CHAPSO/DLPC lipid bicelles (Wang et al., 1998). Both samples contained 2.5 mM protein, 10 mM phosphate buffer at pH 6.5, 0.03% NaN_3, 0.1 mM Pefabloc SC, 600 mM NaCl, and 500 µl of 95% H_2O/5% D_2O in an 5 mm NMR tube.

The cyclophilin A sample was approximately 0.7 mM in 100 mM potassium phosphate buffer at pH 6.5 and 0.03% NaN_3. Solutions of 250 µl (95% H_2O/5% D_2O) were measured in Shigemi microcell tubes. Alignment was achieved by CHAPSO/DLPC/CTAB bicelles (5% total lipid conc.: 1: 5: 0.1; Losonczi and Prestegard, 1998).

All measurement were carried out on Bruker DRX-600 or Bruker DRX-800 (Bruker, Rheinstetten, Germany) spectrometers equipped with standard 5 mm triple-resonance, z-gradient probes. The temperature for all measurements was 303 K. The measurements of the $^1J_{NH}$ splitting in the ^{15}N dimension of 2D ^1H-^{15}N HSQC spectra were collected with 512 (t_1) $\times$ 2048 (t_2) complex data points. $^1J_{NH}$ modulated spectra were collected with 128 (t_1) $\times$ 2048 (t_2) complex data points. Data processing and analysis were performed using either XWinNMR 2.6 (Bruker, Karlsruhe, Germany) or Felix98.0 (MSI, San Diego, CA, USA).

Results and discussion

We have applied the program to three different protein structures: For rhodniin (Friedrich et al., 1993; van de Locht et al., 1995) we calculated a theoretical set of dipolar couplings and pseudocontact shifts using an NMR structure (Maurer and Griesinger, personal communication). A 3D homology search is performed on a restricted database of proteins according to procedure (b) of Figure 2. For cyclophilin A we recorded experimental dipolar couplings and procedure (a) of Figure 2 is used to analyze the experimental data. The dipolar couplings are fitted against the known NMR and X-ray structures, and the orientation of the alignment tensor is determined. The third example is the protein *Hgi*CIC which is currently under investigation in our laboratory. This protein contains a helix-turn-helix motif. Using experimentally derived dipolar couplings a 3D homology search on a restricted set of the PDB was performed.

Rhodniin consists of 103 amino acid and contains two similarly folded domains of 45 amino acids connected by a flexible linker of 10 amino acids. A set of $^1D_{NH}$ dipolar couplings and pseudocontact shifts for amide hydrogens was calculated from the known NMR structure of the protein for the N-terminal domain assuming a specific size and orientation of the alignment tensor and a specific position of a paramagnetic center. Only 36 couplings in rigid parts of the domain were used for the following calculations. The eigenvalues were set to be $S_{zz}^{\mathrm{diag}} = 4.58 \times 10^{-4}$, $S_{yy}^{\mathrm{diag}} = -2.96 \times 10^{-4}$ and $S_{xx}^{\mathrm{diag}} = -1.62 \times 10^{-4}$, amounting to a rhombicity of 0.2. This set of dipolar couplings and pseudocontact shifts is used as an ‘experimental’ test set.

Measured dipolar couplings were fitted to the NMR structure of rhodniin, by omitting and including pseudocontact shifts. As expected, the dipolar couplings are reproduced in the first case (Figure 3, Table 1) when pseudocontact shifts were omitted. With a normalized square deviation of $Q = 0.00$ the tensor size and orientation exactly reproduce the predefined values. The Q-value is found to be 0.08 in the second case when the tensor and the position of the paramagnetic center were recalculated. The paramagnetic center is found with a deviation of 0.786 Å to its original position. This deviation is caused by the grid search step size of 0.5 Å yielding a maximum deviation of $\frac{1}{2}\sqrt{3}$ Å $\approx$ 0.866 Å. This deviation is also the reason for $Q > 0.00$. Additionally, deletion of one, two or three amino acids after residues 15 and 39, as well as the addition of amino acids at the same positions do not influence the result of the calculation. Sequence alignment is always found correctly, irrespective of the usage of pseudocontact shifts.

The ‘experimental’ set of dipolar couplings was fitted to the X-ray structure of ovomucoid (a homologous protein to the N-terminal domain of rhodniin). The ‘experimental’ values as well as the values calculated for the best fit are given in Figure 3, together with the visualization of both structures in the frame of the resulting alignment tensor. The program finds an eight amino acid shift in the sequence alignment (Table 1) which agrees with the primary sequence alignment for rhodniin and ovomucoid. In this case, the normalized square deviation was found to be $Q = 0.30$.

To speed up the process of three dimensional homology search, a subset of 125 folds was extracted from the PDB with a diverse set of folds according to Rost and Sander (1994). Loading the data and calculating secondary structure elements for all proteins in the fold database takes about 5 min on a 450 MHz Pentium II processor. The search itself takes only below 1 s for the whole database, if no gaps are introduced. This time increases to be 48 s if disconnecting of protein parts as explained above with a gap size of up to 5 amino acids is allowed.

The search over this database using the earlier mentioned theoretical set of dipolar couplings for the N terminal domain of rhodniin (a typical Kazal inhibitor) yields ovomucoid (1ovo_a) as 2nd best hit with a Q-value of 0.45 and porcine pancreatic secretory trypsin inhibitor (1tgs_i) as 16th best hit with a Q-value of 0.53. Both proteins are known as Kazal inhibitors and are homologous to rhodniin. In 9 out of these best 16 examples the α-helix of the rhodniin

Table 1. Results of fitting the experimental set of dipolar couplings of the N-terminal domain of rhodniin to ovomucoid. Identical amino acids in both sequences are labeled by | and similar amino acids are labeled by *

rhodniin	:	12	L	H	R	V	C	G	S	D	G	E	T	Y	S	N	P	C	T	L	N	C	A	K	F	N	G	K	P	E	L	V	L	V	H	D	G	C	47
				*		*	|	|	|	|			|	|		|	|	|					|					*	|		|		|		|		|	|	
ovomucoid	:	20	T	R	P	L	C	G	S	D	N	K	T	Y	G	N	P	C	N	F	C	N	A	V	V	E	S	N	P	T	L	T	L	S	H	F	G	C	55

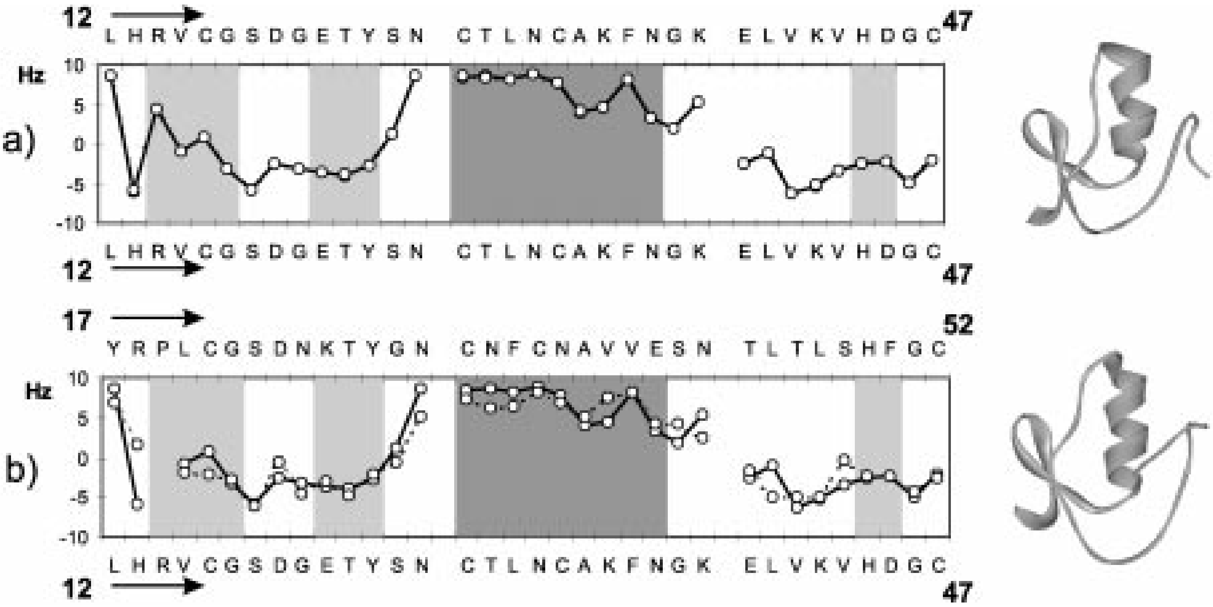

Figure 3. Results for fitting a theoretical set of dipolar couplings for the N-terminal domain of rhodniin to the rhodniin structure (protein **A**) itself (a) and to ovomucoid (protein **B**, a Kazal inhibitor), which is homologous in sequence and structure (b). The black lines indicate the theoretical calculated coupling values, the dotted lines indicate dipolar couplings calculated for the final fit. On the upper x-axis the amino acid number of protein **B**, on the lower x-axis the amino acid number of protein **A** is found. Secondary structure elements are shown by light gray areas (β-sheet) and dark gray areas (α-helix). The three dimensional structures are given in the coordinate system of the tensor (*y*- and *z*-axis are in the paper plane, the *x*-axis is perpendicular to the paper plane).

Table 2. Results of fitting the experimental set of dipolar couplings of the N-terminal domain of rhodniin to rhodniin itself and to an ensemble of eight Kazal inhibitors, some of which are in complex with serine proteases. For 1tbq the data of the N-terminal domain are fitted to the homologous C-terminal domain of rhodniin

Protein name	pdb code	Fit range	*Q*
Rhodniin	–	12–47	0.00
Rhodniin in complex with thrombin (Res.: 2.6 Å)	1tbr	12–47	0.27
Rhodniin in complex with thrombin (Res.: 3.1 Å)	1tbq	65–101	0.27
Ovomucoid	1ovo	20–55	0.30
Human pancreatic secretory inhibitor in complex with trypsin	1cgi	20–55	0.30
Procine pancreatic secretory inhibitor in complex with trypsin	1tgs	19–54	0.32
Human pancreatic secretory trypsin inhibitor	1hpt	20–55	0.36
Pig proteinase inhibitor (Kazal type)	1pce	24–59	0.38
Leech-derived inhibitor with procine in complex with trypsin	1ldt	10–45	0.49

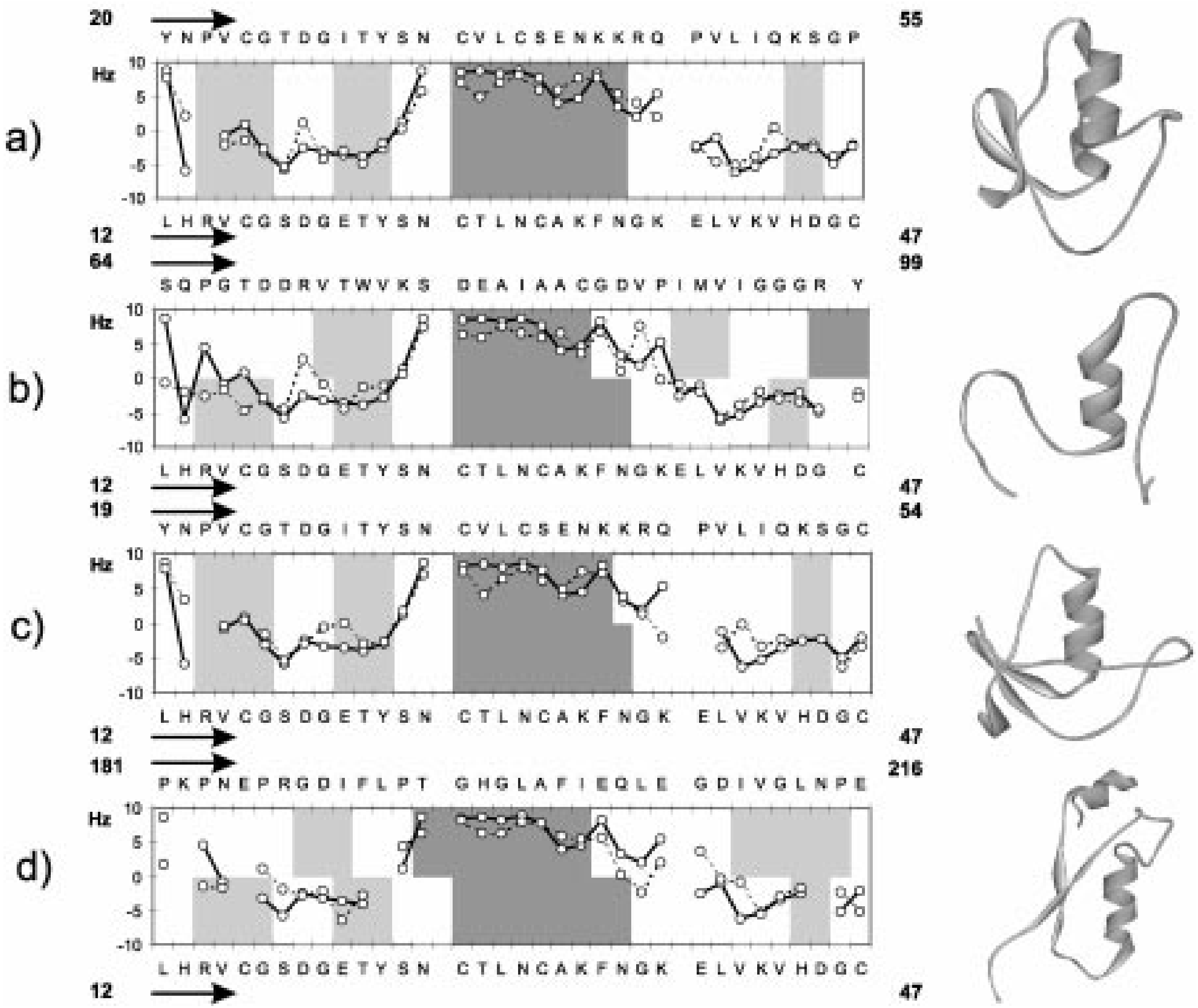

Figure 4. Results for a search in a database of 125 folds extracted from PDB for the theoretical set of dipolar couplings for the N-terminal domain of rhodniin. The black lines indicate the experimental coupling values (protein **A**, rhodniin), dotted lines indicate dipolar couplings calculated for protein **B** from the database. The upper and lower *x*-axes show the amino acid number of protein **B** and protein **A**, respectively. Secondary structure elements are represented similar to Figure 3. The results are ordered by increasing normalized square deviations (*Q*-values). (a) ovomucoid (1ovo_a) with a *Q*-value of 0.30, (b) fragment of an oxidoreductase (6fdr residues 64–99) with a *Q*-value of 0.31, (c) is again a proteinase inhibitor (1tgs_i) with a *Q*-value of 0.32 and (d) is a part of an intramolecular oxidoreductase (4xia_a residues 181–216) with a *Q*-value of 0.35. Subsequent hits have considerably worse matches with *Q*-values above 0.40.

domain is fitted over a β-strand of the protein from the PDB. This observation can be explained by the parallel orientation of N-H^N bond vectors in both secondary structure elements. Dipolar couplings are therefore of the same size in both secondary structure elements which makes a distinction difficult.

Much more significant results with less false positive answers and lower *Q*-values are obtained when secondary structure information from CSI is utilized by two simple rules: first, the alignment of β-strands over α-helices is excluded and second, only residues in well defined secondary structure regions are used for the calculation of *Q*-values. Using these rules, the two Kazal inhibitors of our database are ranked 1st (ovomucoid, 1ovo_a, $Q = 0.30$) and 3rd (porcine pancreatic secretory trypsin inhibitor, 1tgs_i, $Q = 0.32$). Figure 4 presents the first four hits of this search for which structures are displayed in the coordinate system of the tensor. The 2nd result is part of dihydrofolate reductase (6dfr) with a *Q*-value of 0.31 and the fourth result is part of D-xylose isomerase (4xia_a) with a *Q*-value of 0.35. Results (b) and (d) have a

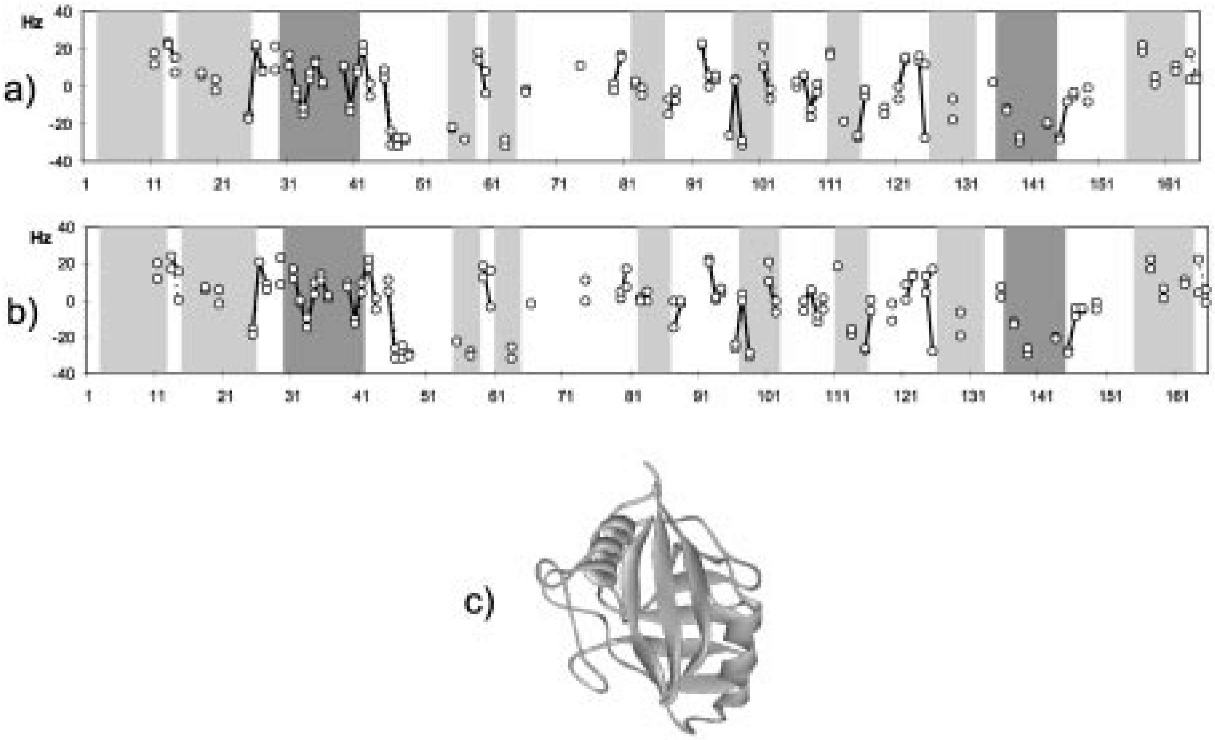

Figure 5. Results of fitting an experimental set of dipolar couplings for cylophilin A (protein **A**) to the NMR structure (a) and to the X-ray structure (b) (protein **B**). Definition of lines and shaded areas is like in Figure 3. *Q*-values are 0.28 and 0.21 for (a) and (b) respectively. The three dimensional structure is given in the coordinate frame of the tensor extracted from the fitting procedure (c).

similarly oriented α-helix and at least one of the three β-strands present in rhodniin with a similar orientation with respect to each other. Since the three β-strands are very short (three residues per strand) and nearly parallel, all dipolar couplings within them are of the same size. Therefore this matches very well with one larger β-strand or an extended region when all N-H^N bonds are parallel (d). Matches (b) and (d) have a primary sequence homology of only 11% and 5%. Thus the program finds 3D homology irrespective of sequence homology.

The result of this first homology search suggests rhodniin to be homologous to other Kazal inhibitors. Therefore a more thorough search for Kazal type inhibitors was performed in the PDB and a subset of such inhibitors was extracted. The *Q*-values of all eight structures range from 0.27 to 0.49 (Table 2).

The second example is cyclophilin for which only 69 fast and easily determinable dipolar couplings were extracted and fitted to the NMR structure (Ottiger et al., 1997) and X-ray (Weber et al., 1982) structures. Results are given in Figure 5 together with the three dimensional structures in the alignment tensor frame of reference. *Q*-values are 0.28 and 0.21 for NMR- and X-ray-structure, respectively. The good agreement of both structures with the experimental data proves that it is not necessary to determine all couplings for fitting. Moreover, the possibility of calculating dipolar couplings for other residues allows to accelerate further interpretation of spectra. While we detect 3D homology to other known cyclophilins, searching in a data bank of 125 folds only finds small parts of the whole sequence, in particular helix-strand-strand motives. It appears that cyclophilin has a rather unique 3D fold.

In the soilbacteria *Herpetosiphon giganteus* many restriction modification systems could be characterized. One of these systems is the *Hgi*CI system of which the C-protein (Controll protein) *Hgi*CIC (expressed with a His_6 tag and a C46 → S mutation) of 10 kDa molecular mass is currently under investigation in our laboratory and was used as a test system for *DipoCoup*. A total of 62 $^1D_{NH}$ dipolar couplings could be extracted for the 88 residue protein *Hgi*CIC. The dipolar couplings range from −7.5 to 7.1 Hz. To establish weak alignment we used CHAPSO/DLPC (1:5) bicelles with a total lipid concentration of 5%.

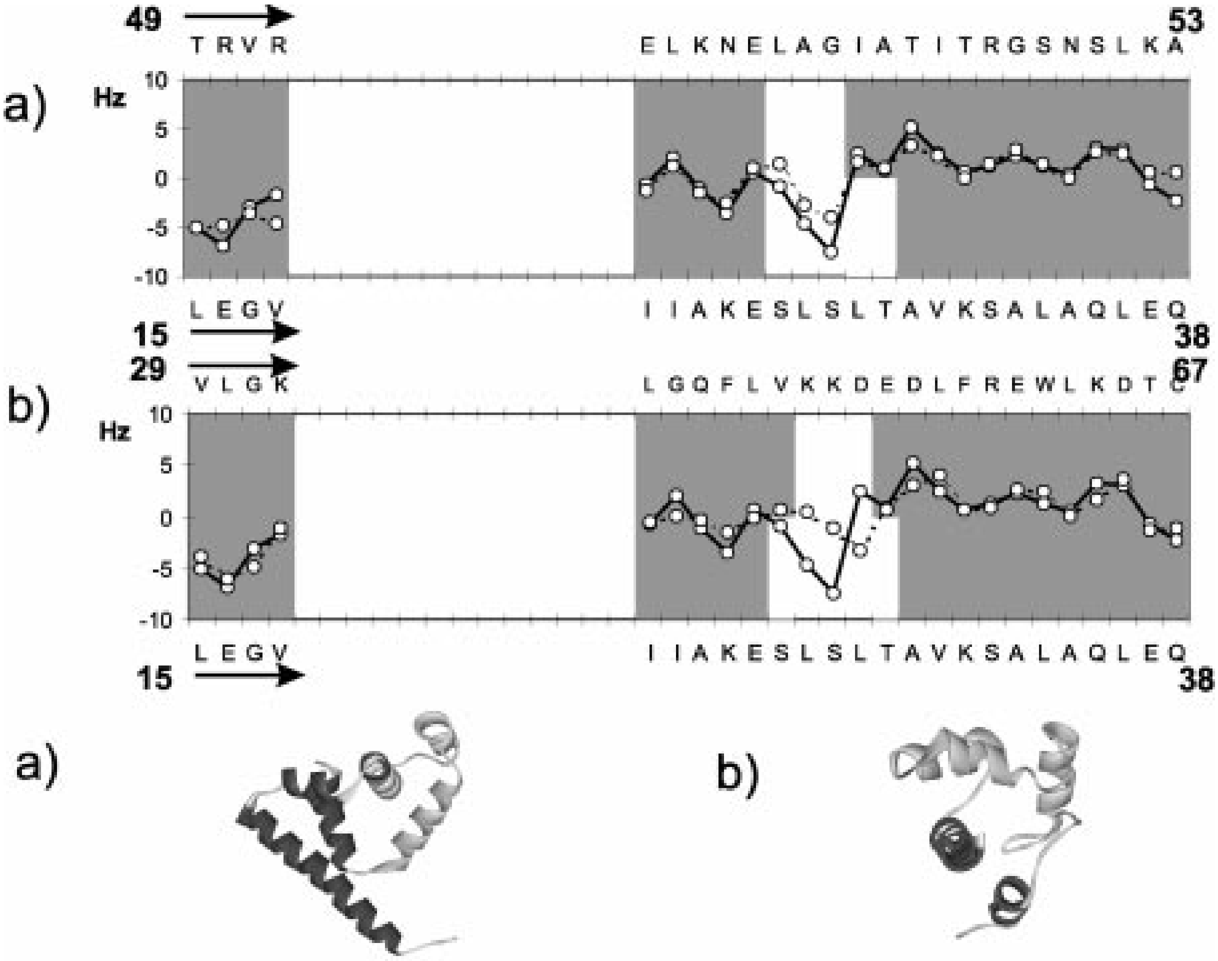

Figure 6. Result of the alignment of residues 15 to 53 of the protein *Hgi*CIC (C46S) to a database especially designed for helix-turn-helix proteins. The results with lowest *Q*-value are trp Repressor (a, $Q = 0.42$) and BAF (b, $Q = 0.43$). Definition of lines and shaded areas is as in Figure 3. The upper and lower x-axes show the amino acid number of protein **B** and protein **A**, respectively. A comparison of the dipolar couplings and corresponding structures in the alignment tensor coordinate system is shown for the two best fits. Light gray parts represent the fitted parts and dark gray parts are not fitted.

Secondary structure alignment indicated the protein might be a typical representative of the helix-turn-helix (HTH) fold family (Brennan and Matthews, 1989; Patto and Sauer, 1992; Harrison, 1999). Therefore we searched for known representatives of the HTH family in the DPInteract (http://arep.med.harvard.edu/dpinteract) database. There are two groups of known HTH proteins. One comprises all α-helical proteins and the other α+β proteins with a HTH motive. With this information we built a database with 19 helix-turn-helix proteins (also at http://krypton.org.chemie.uni-frankfurt.de/∼mj/software.html). From the experimental $^{1}D_{NH}$ dipolar couplings of *Hgi*CIC the alignment tensor was calculated and the alignment search according to Figure 2b was performed with *DipoCoup*, including CSI data. The whole *Hgi*CIC (C46 → S) protein proved to be too large for an alignment with the structures of the database. We therefore partitioned the protein into two overlapping parts. The first part contained the residues 15 to 53 and the second one residues 32 to 71. Both regions can be aligned with parts of proteins in the HTH database. Alignment of the first part (residues 15 to 53) shows good match with the trp repressor (*Q*-value: 0.42) and the cellular factor BAF (*Q*-value of 0.43, Figure 6). The second part (residues 32–71), which includes also the HTH motif, does not match as well as the first stretch of amino acids. We find a best match with the structures of LexA (*Q*-value: 0.64) and with the struc-

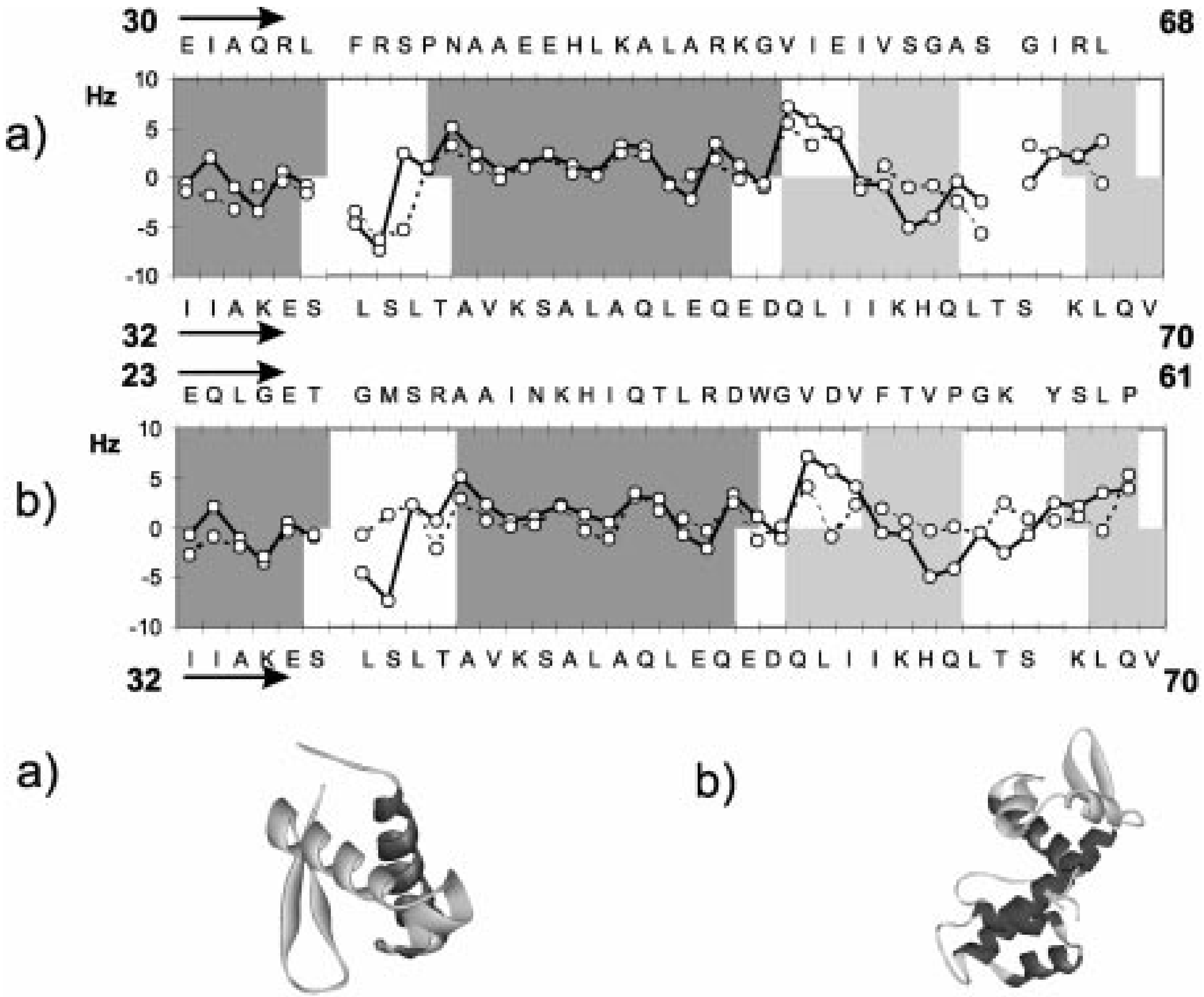

Figure 7. The best match of residues 32 to 71 of the protein *Hgi*CIC (C46S) with HTH protein database is shown. Measured dipolar couplings of the protein are plotted against the calculated dipolar couplings of the two best fitting proteins LexA, and diphtheria toxin repressor (b, $Q = 0.71$). The upper and lower *x*-axes show the amino acid number of protein **B** and protein **A**, respectively. The residue by residue match of the dipolar couplings is much better than the rather high *Q* value would suggest. Light gray parts represent the fitted parts and dark gray parts are not fitted.

ture of the diphtheria toxin repressor (*Q*-value: 0.71). Even though the *Q*-values for the alignment are quite high, the experimental and the calculated dipolar couplings match rather well on a residue by residue basis (Figure 7). A few large deviations can cause large *Q*-values, since *Q* depends quadratically on the deviation of dipolar couplings. Although *HgI*CIC is not very similar to any of the already known HTH-proteins in total, two parts of its structure match known protein folds from which a 3D model of the protein can be derived.

C proteins are also known to bind DNA. The HTH motif in *Hgi*CIC is consistent with the finding that *Hgi*CIC binds DNA as observed by band shift assays. The observed shifts upon DNA titration are most prominent for the amino acids in the helix-turn-helix motif.

Conclusions

We have demonstrated the possibility to use residual dipolar couplings and pseudocontact shifts together with secondary structure information to perform 3D structure homology searches in representative sub-databases of the PDB. We present the program *DipoCoup* which performs this homology search in a fast, accurate and user friendly way. Moreover, *DipoCoup* can be used to perform additional analysis of experimentally determined orientation data or 3D structures of proteins. The program is free for academic use, and can be downloaded from http://krypton.org.chemie.uni-frankfurt.de/∼mj/software.html.

Acknowledgements

This work was supported by the Fonds der Chemischen Industrie. It was further supported by the DFG through (Gr1211/4-4) and the MPG. We thank J.A. Losonczi for making her paper available for us prior to publication. We are grateful to M. Maurer for the NMR structure of rhodniin and for extracting the dipolar couplings of cyclophilin. We are thankful for enlightening discussions with Dr. K. Fiebig. Moreover, we thank A. Zeidler, M. Müller, B. Coligaev and F. Schmäschke for downloading, analyzing and preparing the subset of the PDB. We thank Prof. Dr. G. Fischer and Dr. H. Bang (MPI, Halle) for the preparation of cyclophilin A and Prof. Dr. M. Kröger and Dr. K. Meyer-Rogge (University Gießen) for the preparation of *Hgi*CIC (His$_6$) [C46 → S]. All spectra were recorded at the Large Scale Facility for Biomolecular NMR at the University of Frankfurt. W.P. and J.M. are both supported by a Kekulé stipend of the Fonds der Chemischen Industrie.

References

Annila, A., Aitio, H., Thulin, E. and Drakenberg, T. (1999) *J. Biomol. NMR*, **14**, 223–230.

Bax, A. and Ottiger, M. (1998) *J. Am. Chem. Soc.*, **120**, 12334–12341.

Bax, A. and Tjandra, N. (1997) *J. Biomol. NMR*, **10**, 289–292.

Brennan, R.G. and Matthews, B.W. (1989) *J. Biol. Chem.*, **264**, 1903–1906.

Clore, G.M. and Garrett, D.S. (1999) *J. Am. Chem. Soc.*, **121**, 9008–9012.

Clore, G.M., Gronenborn, A.M. and Bax, A. (1998) *J. Magn. Reson.*, **133**, 216–221.

Clore, G.M., Gronenborn, A.M. and Tjandra, N. (1998) *J. Magn. Reson.*, **131**, 159–162.

Cornilescu, G., Marquardt, J.L., Ottiger, M. and Bax, A. (1998) *J. Am. Chem. Soc.*, **120**, 6836–6837.

Delaglio, F., Kontaxis, G. and Bax, A. (2000) *J. Am. Chem. Soc.*, **122**, 2142–2143.

Fischer, D. and Eisenberg, D. (1999) *Curr. Opin. Struct. Biol.*, **9**, 208–211.

Fischer, M.W.F., Losonczi, J.A., Weaver, J.L. and Prestegard, J.H. (1999) *Biochemistry*, **38**, 9013–9022.

Friedrich, T., Kröger, B., Bialojan, S., Lemaire, H.G., Höffken, H.W., Reuschenbach, P., Otte, M. and Dodt, J. (1993) *J. Biol. Chem.*, **268**, 16216–16220.

Ghose, R. and Prestegard, J.H. (1997) *J. Magn. Reson.*, **128**, 138–143.

Harrison, S.C. (1991) *Nature*, **353**, 715–719.

Losonczi, J.A., Andrec, M., Fischer, M.W.F. and Prestegard, J.H. (1999) *J. Magn. Reson.*, **138**, 334–342.

Losonczi, J.A. and Prestegard, J.H. (1998) *J. Biomol. NMR*, **12**, 447–451.

Meiler, J., Blomberg, N., Nilges, M. and Griesinger, C. (2000) *J. Biomol. NMR*, **16**, 245–252.

Moult, J. (1999) *Curr. Opin. Biotechnol.*, **10**, 583–588.

Ojennus, D.D., Mitton-Fry, R.M. and Wuttke, D.S. (1999) *J. Biomol. NMR*, **14**, 175–179.

Ottiger, M., Zerbe, O., Güntert, P. and Wüthrich, K. (1997) *J. Mol. Biol.*, **272**, 64–81.

Patto, C.O. and Saurer, R.T. (1992) *Annu. Rev. Biochem.*, **61**, 1053–1095.

Peti, W. and Griesinger, C. (2000), *J. Am. Chem. Soc.*, **122**, 3975–3976.

Rost, B. and Sander, C. (1994) *Proteins Struct. Funct. Genet.*, **19**, 55–72.

Sali, A. (1998) *Nat. Struct. Biol.*, **5**, 1029–1032.

Saupe, A. (1968) *Angew. Chem. Int. Ed. Engl.*, **7**, 97–102.

Spera, S. and Bax, A. (1991) *J. Am. Chem. Soc.*, **113**, 5490–5492.

Tjandra, N. and Bax, A. (1997) *Science*, **278**, 1111–1113.

Tjandra, N., Grzesiek, S. and Bax, A. (1996) *J. Am. Chem. Soc.*, **118**, 6264–6272.

Tolman, J.R., Flanagan, J.M., Kennedy, M.A. and Prestegard, J.H. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 9279–9283.

van de Locht, A., Lambda, D., Bauer, M., Hubert, R., Friedrich, T., Kroeger, B., Hoffken, W. and Bode, W. (1995) *EMBO J.*, **14**, 5149–5155.

Wang, H., Eberstadt, M., Olejniczak, E.T., Meadows, R.P. and Fesik, S.W. (1998) *J. Biomol. NMR*, **12**, 443–446.

Weber, E., Papamokos, E., Bode, W., Huber, R., Kato, I. and Laskowski, M. (1982) *J. Mol. Biol.*, **158**, 515–520.

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. and Sykes, B.D. (1994) *J. Biomol. NMR*, **4**, 171–180.

Wishart, D.S., Sykes, B.D. and Richards, F.M. (1992) *Biochemistry*, **31**, 1647–1651.