J-Bio NMR 361

Calculation of symmetric multimer structures from NMR data using a priori knowledge of the monomer structure, co-monomer restraints, and interface mapping: The case of leucine zippers

Sean I. O'Donoghue[a,]*, Glenn F. King[b] and Michael Nilges[a]

[a]*European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69012 Heidelberg, Germany*
[b]*Department of Biochemistry, University of Sydney, Sydney, NSW 2006, Australia*

Received 2 January 1996
Accepted 16 April 1996

Keywords: Nuclear Overhauser enhancement; Symmetric multimers; Solution structure; Leucine zippers

Summary

NMR studies of symmetric multimers are problematic due to the difficulty in distinguishing between intra-, inter-, and co-monomer (mixed) NOE signals. Previously, one of us described a general calculation strategy called dynamic assignment by which this difficulty can be overcome [Nilges, M. (1993) *Proteins*, **17**, 297 309]. Here we describe extensions to the method for handling many co-monomer NOEs and for taking advantage of prior knowledge of monomer structures. The new protocol was developed for the particularly difficult case of leucine zipper (LZ) homodimers, for which the previous protocol proved inefficient. In addition to the problem of dimer symmetry, LZs have a particularly high proportion of co-monomer NOE signals and a high degree of repetition in sequence and structure, leading to significant spectral overlap. Furthermore, the leucine zipper is a rather extended (as opposed to globular) protein domain; accurately determining such a structure based only on the very short distances obtainable by NMR is clearly a challenge to the NMR structure determination method. We have previously shown that, for LZ homodimers, many of the backbone–backbone NOESY cross peaks can be unambiguously assigned as intra-monomer, enabling approximate monomer structures to be calculated. Using model and experimental data sets, we verified that the new protocol converges to the correct dimer structure. The results show that short-range NMR distance data can be sufficient to define accurately the extended LZ. The protocol has been used to derive a novel solution structure of the c-Jun LZ domain. Based on these calculations, we propose the protocol as a prototype for the general case of symmetric multimers where the monomer structure is known.

Introduction

Quite often in protein ^{1}H NMR spectra, nuclei from different hydrogen atoms on the same polypeptide chain have indistinguishable resonant frequencies; we refer to this as dispersion degeneracy. NOE cross peaks which occur at degenerate frequencies can be converted to ambiguous distance restraints (Nilges, 1993). In some cases (methyl, methylene, or ring protons), the problem of ambiguity can be overcome using pseudoatoms (Wüthrich et al., 1983) or $\langle r^{-6} \rangle$ averaging methods (Brünger et al., 1986). In general, however, conventional structure calculation methods cannot use ambiguous restraints directly. Hence, the remaining ambiguous restraints are normally excluded from initial structure calculations; some are assigned later using iterative strategies.

Proteins often occur as symmetric multimers; in such cases, we cannot afford to ignore ambiguous data. In a perfectly symmetric multimer, all symmetry-related hydrogens will have equivalent magnetic environments, hence

*To whom correspondence should be addressed. All the protocols in this paper are freely available (http://www.nmr.embl-heidelberg.de/nilges/).

Abbreviations: 3D, three-dimensional; GCN4-c, crystal structure of the GCN4 LZ homodimer; GCN4-s, solution structures of GCN4; GSYM, global symmetry; Jun-m, model structure of the Jun LZ homodimer; Jun-s, solution structure of Jun; LZ, leucine zipper; MFP, mean force potential; MDSA, molecular dynamical simulated annealing; NCS, noncrystallographic symmetry; NOE, nuclear Overhauser enhancement; rmsd, root-mean-square deviation; vdW, van der Waals.

they will be degenerate (i.e., only one monomer is 'seen' in NMR spectra); we refer to this as symmetry degeneracy. Dispersion degeneracy and symmetry degeneracy are quite distinct: while dispersion degeneracy can in principle be improved with more highly resolved spectra or better acquisition, symmetry degeneracy cannot. In NOESY spectra of multimers, each NOE cross peak falls into one of three classes: (i) '*intra*-monomer', those arising solely from dipolar coupling between protons within the same monomer; (ii) '*inter*-monomer', those arising solely from coupling between protons on different monomers; and (iii) we introduce the term '*co*-monomer' to describe NOEs which arise from *both* couplings between protons on the same monomer *and* couplings between protons on different monomers. In the case of symmetric multimers, due to symmetry degeneracy it is not possible a priori to distinguish between intra-, inter-, and co-monomer NOE signals, so in principle *every* distance restraint is ambiguous.

Two experimental approaches to the symmetry degeneracy problem have been tried. In the first method, deuterated analogues of the protein are made and mixed with the protonated protein; comparing cross-peak intensities with those from spectra of the fully protonated protein enables some discrimination between intra- and inter-monomer NOEs (Arrowsmith et al., 1991). This method works for dimers, but the analysis becomes intractable for most higher-order multimers. The second method uses mixtures of unlabelled and ^{13}C,^{15}N-labelled protein for heteronuclear filtered and separated 2D NMR experiments (Folkers et al., 1993; Folmer et al., 1995). This method also has problems for higher-order multimers: not all ambiguities can be resolved. This was highlighted recently by problems encountered in determining the solution structure of the p53 homotetramer (Clore et al., 1995; Lee et al., 1995).

We have recently suggested a computation method, called 'dynamic assignment', which handles both dispersion (Nilges, 1995) and symmetry ambiguity (Nilges, 1993; O'Donoghue et al., 1993). It has several advantages: it does not require production of labelled protein; all information in the spectra can be used to direct the structure calculation; and it generalises to any kind of symmetry. In many cases, the best approach would be to combine both methods. The dynamic assignment method is similar to the $\langle r^{-6} \rangle$ averaging method for degenerate methyl protons (Levy et al., 1989), although there is a subtle difference (Nilges, 1993). Nilges previously described a protocol which uses the dynamic assignment method to calculate structures in the general case of completely symmetric multimers (1993). The protocol was tested on three symmetric dimer structures; in each case, the protocol converged to the correct dimer structure.

In this paper we have applied the dynamic assignment method to the leucine zipper (LZ) domain of bZIP homodimers. In addition to their fundamental biological role in eucaryotic transcription activation, these proteins have been implicated in the oncogenic transformation of cells. There have been several previous NMR studies of leucine zippers, namely the homodimers of GCN4 (Oas et al., 1990; Saudek et al., 1990,1991), cyclic GMP-dependent protein kinase (Atkinson et al., 1991), and c-Jun (Junius et al., 1993). To date, however, all have failed to calculate a complete dimer structure from the NMR data. Our initial calculations using the previous protocol also gave a very low convergence rate. Since dynamic assignment puts an additional computational load on the calculation, it is important to have a reasonably high convergence rate.

The LZ domains are a particularly difficult case for four reasons in addition to the symmetry degeneracy. Firstly, the dispersion degeneracy problem is severe in the case of the LZs due to the high degree of repetition in the sequences and structure of LZs. Secondly, since the symmetry axis coincides with the principal moment of the molecule, we would expect fewer inter-monomer NOEs (which would strongly drive the calculation towards the correct structure), and more co-monomer NOEs (weaker restraints which are more difficult to assign). Many inter-monomer NOEs will be between symmetry-related hydrogens and hence cannot be measured in a homonuclear experiment, since they will occur on the diagonal of the NOE spectrum. Thirdly, since the overall shape is highly extended – probably the most extended (as opposed to globular) protein domain studied to date – we expect to obtain less NOE connectivities per residue. Since the number of NOEs per residue is a basic guide to the quality of an NMR structure, we are likely to have difficulties producing accurate structures. Finally, since NMR structure determination is based on the very short distances (< 5 Å) measurable from NOE spectra, it is indeed an open question whether or not the method can define the global structure (coiled-coil pitch) of such an extended protein.

Given the assumption that the structure is a coiled coil, but without making any assumptions about which residues form the interface, many backbone–backbone NOEs can be unambiguously assigned as intra-monomer (O'Donoghue et al., 1993). Junius et al. (1993) recently used this method to calculate an approximate monomer structure for the Jun homodimer. As previously argued (O'Donoghue et al., 1993), we believe this is a better approach than those used to calculate monomer structures for GCN4 (Saudek et al., 1990,1991) and cGMP-dependent protein kinase (Atkinson et al., 1991). Given that we have a reasonably accurate prior knowledge of the monomer structure, we develop a new protocol using the dynamic assignment method to exploit this knowledge. Using both model and experimental distance data sets, we show that the new protocol has a much higher convergence rate towards the correct LZ structure than the previous protocol. The results show that NMR can indeed accurately define such extended structures. We have used this protocol to derive a novel solution structure for the c-Jun LZ homodimer. This is the first complete solution structure of an LZ domain (Junius et al., 1996).

Methods

General calculation procedures

We did all calculations, except where noted otherwise, in X-PLOR 3.1 (Brünger, 1992) using the simplified force field given by the parallhdg.pro and topallhdg.pro files; this field uses uniform values for the energy constants of each geometric energy term: $k_{bond} = 1000$ kcal mol^{-1} Å^{-2} for all bonds, $k_{angle} = 500$ kcal mol^{-1} rad^{-2} for all angles, and $k_{planar} = 500$ kcal mol^{-1} rad^{-2} for all dihedral-angle restraints which maintain planarity and chirality.

Nonbonded interactions were calculated using the 'repel' potential (Nilges et al., 1988b):

$$E_{vdW} = k_{vdW} \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} \left[\mathrm{Max}\left(0, \left[s\left(r_i + r_j\right)\right]^2 - d_{ij}^2\right)\right]^2 \qquad (1)$$

where $k_{vdW} = 4$ kcal mol^{-1} Å^{-4}; A is the total number of atoms in the multimer; the scale factor s is usually set to 0.8 giving the atomic radii used in DISGEO (Havel and Wüthrich, 1984); r_i and r_j are the van der Waals (vdW) radii of atoms i and j, respectively; and d_{ij} is the distance between atoms i and j.

In this paper, we regard each energy constant, for example k_{NOE}, as having a constant value (in this case 50 kcal mol^{-1} Å^{-2}) at all stages in the refinement protocols. However, the contributions of the energy terms to the total energy, e.g. $w_{NOE}E_{NOE}$, are varied during the protocol by changing the dimensionless weight factors, w_{NOE}. (In some cases we used separate weights for the ambiguous and unambiguous NOE restraints, i.e. $E_{NOE} = w_{ambig}\,E_{ambig} + w_{unambig}\,E_{unambig}$.) This convention makes it meaningful to compare energy values at different stages of the refinement. Similarly, for dihedral-angle restraints arising from coupling constants we used the usual square-well potential with an energy constant of $k_{CDIH} = 200$ kcal/mol, but varied the weight w_{CDIH}.

We also use a similar convention for the other experimental energy terms: $k_{NCS} = 2$ kcal mol^{-1} Å^{-2}, $k_{GSYM} = 0.5$ kcal mol^{-1} Å^{-2}, $k_{HEPTAD} = 2$ kcal mol^{-1} Å^{-2} (see below for an explanation of these terms).

Since LZs are rather elongated molecules with an unusually large radius of gyration, we used the size-independent ρ-factor (Maiorov and Crippen, 1995) as a measure for comparing similarity of structures, as well as the conventional root-mean-square deviation (rmsd).

The dynamic assignment method

Here we give a mathematical description of the dynamic assignment method for the general case of a completely symmetric multimer of M monomers.

Let $V^P = \{V_n^P\!:\! n = 1,\ldots,N\}$ denote the set of N cross-peak volumes obtained from the NOE spectra of a particular protein P. We denote the frequencies corresponding to these cross peaks as $F1_n$ and $F2_n$; the sets of atoms which are assigned to these frequencies are denoted A_n and B_n, respectively; we partition these sets into subsets for each monomer, so that $A_{n,m} = \{a_{n,m,i}\!:\! i = 1,\ldots,Ea_n\}$ denotes the set of Ea_n protons on the mth monomer with equivalent frequency of a particular value; $B_{n,m}$ is defined similarly. Usually, the number of equivalent protons on each monomer, Ea and Eb, will be 1, 2 (for some methylene groups and aromatic protons), or 3 (methyl groups); sometimes it will be higher due to dispersion degeneracy. In the following, we drop the last subscript for Ea and Eb values of 1, e.g. $A_{n,m} = \{a_{n,m}\}$. In theory, the cross-peak volume V_n is due to the sum of interactions between each A_n atom and each B_n atom. Assuming the isolated two-spin approximation, the total volume is related to the interproton distances (d) by the following equation*:

$$V_n = c \sum_{\mu=1}^{M} \sum_{m=1}^{M} \sum_{i=1}^{Ea_n} \sum_{j=1}^{Eb_n} d\left(a_{n,\mu,i}, b_{n,m,j}\right)^{-6} \qquad (2)$$

In practice, many of the distances will be greater than 5 Å and their contribution to V_n will be negligible. The scaling factor c is usually calculated once for each spectrum (some calibration methods use several different factors for different classes of NOEs). When dealing with symmetric multimers, this calibration can be problematic. To calculate c, we chose one peak, V_c, which we know arises from two clearly resolved protons (i.e., $Ea_c = Eb_c = 1$) with a known interatomic distance, e.g. two protons in a resolved methylene group, or certain pairs of backbone protons within known secondary structure elements. The problem in the symmetric multimer case is that we need to consider the interactions with equivalent protons on other monomers; strictly speaking, we can only use such a pair of protons, a_c and b_c, for calibration if we know that

$$d(a_{c,1}, b_{c,1})^{-6} \gg d(a_{c,1}, b_{c,m})^{-6} \qquad (3)$$

for all m ≠ 1. Most methylene groups will satisfy this condition. In the case of the symmetric LZ homodimers, many backbone–backbone cross peaks will also satisfy this condition (O'Donoghue et al., 1993). Assuming this condition, we can calculate c from Eq. 2:

$$c = V_c\, d(a_{c,1}, b_{c,1})^6 / M \qquad (4)$$

In practice, due to uncertainty introduced by possible non-zero inter-monomer interactions, it is best to calcu-

*In the case of degenerate methyl groups, the exponent should be −3 rather than −6 to account for the rapid motional averaging of the hydrogens. In practice, this difference is negligible.

late c from several reference distances. The effect of intermonomer terms on Eq. 4 would be to increase c artefactually, so unusually large values should be ignored.

Having determined c, we can then convert all the observed volumes into restraint distances (D_n) using the following equation:

$$D_n = (V_n/cM)^{-1/6} \qquad (5)$$

The division by M ensures consistency with the calibration distances, i.e., $D_c = d(a_{c,1},b_{c,1})$. During refinement, each restraint distance is compared with the d^{-6} sum of all distances in the model structure which may contribute to the restraint, viz.:

$$\overline{D}_n = \left(\frac{1}{M} \sum_{\mu} \sum_{m} \sum_{i} \sum_{j} d(a_{n,\mu,i}, b_{n,m,j})^{-6} \right)^{-1/6} \qquad (6)$$

The summed distance $\overline{D}_n$ is used exactly as a standard distance restraint originating from an unambiguous NOE. During refinement, the structure is constrained to satisfy the experimental distance restraints using the 'soft' potential function (Nilges et al., 1988c) which switches between flat, square, and asymptotic behaviour:

$$E_{NOE} = k_{NOE} \sum_{n} \begin{cases} 0 & ; \overline{D}_n < D_n \\ (\overline{D}_n - D_n)^2 & ; \sigma > \overline{D}_n \geq D_n \\ \alpha(\overline{D}_n - \sigma)^{-1} + \beta(\overline{D}_n - \sigma) + \chi & ; \overline{D}_n \geq \sigma \end{cases} \qquad (7)$$

where the parameters α and χ are determined by the requirement that the function is continuous and differentiable at the switching distance σ, and β is a settable parameter. This potential form can be used for both dispersion and symmetry-ambiguous restraints. Thus, we restrict the search space during refinement to conformations which satisfy the ambiguous restraints. Finally, by looking at the convergence of the ensemble of final structures, we can decide which protons contribute significantly to each restraint, and hence determine the assignments.

Co-monomer restraints

Here, we describe an extension of the dynamic assignment method for dealing with co-monomer NOEs. Consider a NOE cross peak arising between atoms $\{a_{n,1},a_{n,2}\}$ (a pair of symmetry-related atoms on monomers 1 and 2, respectively) and atoms $\{b_{n,1},b_{n,2}\}$ in the spectra of a symmetric dimer. From Eq. 2, the total volume of the cross peak, V_n, is related to the intra- and inter-monomer distances by:

$$V_n/c = 2\, d(a_{n,1},b_{n,1})^{-6} + 2\, d(a_{n,1},b_{n,2})^{-6} \qquad (8)$$

When a and b are distant from the dyad axis, one of these distances will be much longer than the other, hence the inverse sixth power will be negligible compared with the other, i.e. the NOE is either intra-monomer or intermonomer, but does not have significant contribution from both terms. However, when a and b are close to the dyad axis, both terms may contribute significantly to V_n, i.e., we have a co-monomer NOE. In this case, the following two conditions hold:

$$d(a_{n,1},b_{n,1}) \leq d_l \qquad (9)$$

and

$$d(a_{n,1},b_{n,2}) \leq d_l \qquad (10)$$

where d_l is the upper limit distance for a 'significant' NOE signal; this is conventionally taken to be 5 Å.

Now suppose that we have identified a NOE as co-monomer, either from a series of structure calculations, or from selective labelling experiments. In the dynamic assignment method above, each restraint is expressed using only the inverse sixth power sum as in Eq. 6. But this does not ensure that both Eqs. 9 and 10 are satisfied, i.e., that both the inter- and intra-monomer distances are less than d_l. Particularly in cases where there are many co-monomer constraints, as in the LZ proteins, it is important to enforce that the structure satisfies Eq. 10. Thus for each NOE which is assigned as co-monomer, we add two extra restraints specifying that the inter-monomer distances $d(a_{n,1},b_{n,2})$ and $d(a_{n,2},b_{n,1})$ are less than d_l, in addition to the normal summed distance restraint used in dynamic assignment.

Specifying the symmetry

The symmetry of the dimer was enforced using the two-term approach proposed by Nilges for specifying the symmetry of symmetric multimers (Nilges, 1993). One term applies a force which acts to keep the monomers superimposable using the noncrystallographic symmetry (NCS) restraint option in X-PLOR. The second term ensures that the relative orientations of the monomers are symmetric, using the global symmetry (GSYM) potential in X-PLOR 3.1. In this potential, we specify some number, G, of pairs of atoms a_g and b_g and restrain all intermonomer distances between them. For specifying dimer symmetry, we use:

$$E_{GSYM} = K_{GSYM} \sum_{g=1}^{G} [d(a_{g,1},b_{g,2}) - d(a_{g,2},b_{g,1})]^2 \qquad (11)$$

This potential allows the structure to evolve its own axis of symmetry during refinement. Clearly, it is not practical to use all possible combinatorial pairs. Fortunately, with the NCS constraint, it is not necessary; it is sufficient to use only a small subset of pairs, provided that the subset somehow spans all residues of the monomer in an equal manner. But which atom pairs should be chosen? Nilges (1993) previously proposed the following subset: $a_g = g\,C^\alpha$, $b_g = (R - g + 1)C^\alpha$ where $g = 1,...,R$: here, R is the number of residues in each monomer, and r C^α indicates the C^α atom of residue r; we refer to this as systematic selection. While systematic selection gave good results for the three symmetric dimers used previously (Nilges, 1993), in the case of the coiled coils reported here, it proved unsatisfactory (see Results). Instead, we propose a more general selection: the a_g atoms are chosen as before, but the b_g atoms are selected at random from all possible atoms; thus, this is called randomised selection.

Distance-restraint sets

We used two homodimeric LZ structures for constructing model sets of NOE distance restraints. The first structure, denoted GCN4-c, was based on the crystal structure of the LZ domain (residues 249 to 281) of yeast GCN4 (Protein Data Bank deposition code 2ZTA; O'Shea et al., 1991) to which we added hydrogen atoms using the X-PLOR HBUILD facility. The second structure, denoted Jun-m, was an all-atoms model structure of the LZ domain (residues 285 to 323) of human Jun (O'Donoghue et al., 1993). This model was built beginning with α-carbons positions calculated from the coiled-coil equation (Crick, 1953) with a pitch of 181 Å, and a radius of 4.65 Å (taken from the values reported for GCN4-c by O'Shea et al., 1991); the coordinates of the other atoms were calculated using the side-chain-building method described by Nilges and Brünger (1991). The final model is strictly symmetric, and has a coiled-coil phase angle of $\theta = 33°$ (found to give the lowest energy; O'Donoghue et al., 1993).

We generated two model sets of distance restraints (denoted $D^{GCN4\text{-}c}$ and $D^{Jun\text{-}m}$ from GCN4-c and Jun-m, respectively) using the following procedure implemented in X-PLOR. Firstly, we constructed a list of all possible sets of protons, A_n and B_n, which could in theory be resolved and produce a cross peak in a NOESY spectrum. At this stage, equivalent protons on different monomers were always regarded as ambiguous and hence were grouped together. All methyl and methylene protons were assumed to have equivalent frequency and hence these were also grouped together. We also excluded rapidly exchanging protons (the amide protons of the N-terminus and the side chains of asparagine and glutamine; the amine protons of lysine; the indole proton of tryptophan; the imidazole proton of histidine; the guanidinium protons of arginine; the hydroxyl protons of serine, threonine, and tyrosine; and the sulphydryl proton of cysteine). For each line in the remaining list, we calculated a restraint distance using Eq. 5 with $M = 2$. Restraint distances greater than 5.0 Å were excluded. The remaining distances were then binned into three groups (D_n less than 2.7, 3.5, or 5.0 Å) and written into a NOE restraint file in X-PLOR format. These are referred to as initial model distance sets.

Two other distance restraint sets were used in this study, both derived from ^{1}H NMR experiments. The first, denoted $D^{GCN4\text{-}s}$, was the data set obtained by Saudek et al. (1991) for the LZ peptide (residues 247 to 281) of yeast GCN4. We converted $D^{GCN4\text{-}s}$ from DIANA/DISMAN format into X-PLOR format using the fmtoxpupl routine written by Güntert (ETH, Zürich). The 25 hydrogen-bond restraints per monomer obtained by Saudek et al. were included in the calculations involving both $D^{GCN4\text{-}c}$ and $D^{GCN4\text{-}s}$. The spectra obtained by Saudek et al. could not be used to derive any dihedral-angle information, hence we used no dihedral angle restraints in our GCN4 calculations. The other distance-restraint set used, denoted $D^{Jun\text{-}s}$, was derived from NOESY spectra of Jun by Junius et al. (1996); 33 dihedral-angle and 31 hydrogen-bond restraints (per monomer) have been derived for Jun, so these were used in both the $D^{Jun\text{-}m}$ and $D^{Jun\text{-}s}$ calculations. Hydrogen-bond restraints of 2.2 Å were used between O and HN atoms, and 3.3 Å between O and N atoms.

As we have shown previously (O'Donoghue et al., 1993), many distances between α-, β-, and NH-protons in a symmetric coiled coil can be unambiguously assigned as intra-monomer. Hence, we partitioned each distance-restraint set into two subsets for unambiguous or ambiguous distances. Due to increased overlap and other complications in the experimental spectra, the subsets derived from the model sets $D^{GCN4\text{-}c}$ and $D^{Jun\text{-}m}$ were much larger than those obtained from $D^{GCN4\text{-}s}$ and $D^{Jun\text{-}s}$. Thus, we randomly deleted restraints from the initial model subsets until we were left with the same number of distances as in the corresponding experimental subsets (Table 1). This deletion process was repeated for each structure calculation. The resulting model distance sets are denoted $D^{GCN4\text{-}cd}$ and $D^{Jun\text{-}md}$.

The $D^{GCN4\text{-}cd}$ and $D^{Jun\text{-}md}$ sets have roughly the same amount of information as the corresponding experimental sets, although we have not accounted for all limitations and systematic biases which affect the experimental data sets. However, for testing protocols, the model data sets have the advantage that we know exactly what the 'correct' structures should be. Thus convergence to the original structures is an exact measure of the success of the protocols.

Correlation statistics

The number of inter- and co-monomer NOEs, p_r, predicted for each residue, r, in GCN4-c was calculated from $D^{GCN4\text{-}c}$. To compare with the number of remaining ambiguous restraints per residue, a_r, the p_r numbers were scaled by the factor $f = \Sigma a_r/\Sigma p_r$. Then the correlation between (a_r, fp_r) pairs was assessed using Kendall's τ-test

TABLE 1
STATISTICS ON DISTANCE-RESTRAINT SETS

Distance-restraint set	Number of restraints[a]	Ambiguous[b] (%)	Unambiguous[b] (%)	Restraints per residue[c]	Intra-monomer[d] (%)	Inter-monomer[d] (%)	Co-monomer[d] (%)
GCN4-c	1095	67	33	35	83	9	8
GCN4-s	356	78	22	8	80	1	6
Jun-m	1445	64	36	34	85	6	9
Jun-s	1334	84	16	12	39	0.5	2

[a] The number of restraints per monomer. There are several reasons for the low number of restraints for $D^{GCN4\text{-}s}$: the peptide did not give a well-resolved spectrum; many 'structurally irrelevant' unambiguous distances were removed; many restraints had dispersion degeneracy, and these were not included in the restraint list; finally, all weak NOEs were removed to minimise the number of inter-monomer restraints (since this data set was used to calculate monomer structures). In contrast, $D^{Jun\text{-}s}$ was derived from well-resolved spectra and includes 'structurally irrelevant' and ambiguous restraints resulting from dispersion degeneracy.

[b] The initial restraint sets were subdivided into ambiguous and unambiguous subsets using the rules proposed by O'Donoghue et al. (1993).

[c] For $D^{GCN4\text{-}c}$ and $D^{Jun\text{-}m}$, the number of NOEs per residue is calculated from the complete set of all distances < 5 Å in the initial structures, i.e. GCN4-c and Jun-m. For $D^{GCN4\text{-}s}$ and $D^{Jun\text{-}s}$, we count only the number of NOEs unambiguously assigned from the final ensembles GCN4-si and Jun-si.

[d] Gives the percentage of NOEs assigned as either inter-monomer, intra-monomer, or co-monomer. For $D^{GCN4\text{-}c}$ and $D^{Jun\text{-}m}$, the assignments were directly calculated from the initial structures. For $D^{GCN4\text{-}s}$ and $D^{Jun\text{-}s}$, the assignments were calculated from the final structures (GCN4-si and Jun-si), and a significant percentage of the restraints remained ambiguous (hence the percentages do not sum to 100%).

(e.g. see Press et al., 1986). The test calculates the probability that the observed correlation (or better) occurs by chance alone.

Interface filter

In the case of GCN4-s, we were able to identify all residues involved in the interface between the two monomers using the above correlation statistics. Knowledge of the complete set of interface residues in a multimeric structure enables us to design an interface filter which screens out structures that do not have the correct interface. The filter uses the following principle: each interface residue must be in contact with at least one interface residue on a separate monomer. For the current purposes, we define a contact between two interface residues as meaning that the α-carbons are within 9 Å, consistent with the contacts between interface residues observed in GCN4-c. We implemented this filter in X-PLOR by defining an ambiguous distance restraint from each interface residue to all interface residues on the opposite monomer. We did not use this filter as an additional constraint during the structure calculation; rather, we applied the filter to sets of final structures calculated from our dimer protocols. We selected only those structures in which all interface residues satisfy the above restraint. This yields sets of final structures from which assignments can be made.

Naming of protocols

In this work, we developed and compared many alternative protocols. We have devised a scheme for naming these protocols, and also future protocols from our group. There are two purposes for the scheme: to allow textual reference to specific protocols, and to facilitate access via Internet. The protocol name needs to be short enough to be a computer file-name, but should contain enough detail to specify the kind of protocol (molecular dynamical simulated annealing, metric-matrix distance geometry, etc.), the molecular system that it is designed for (single asymmetric molecule, symmetric dimer, multimer, etc.), and the type of initial structure that is expected (random chain, well-defined molecule, refined model, etc.). The names of the protocols mentioned in this paper, and the exact scheme for generating the names are given in Table 2.

Calculation of monomer structures

Monomer structures were generated from the unambiguous distance subsets, together with the hydrogen-bond and dihedral-angle restraints, using the protocols nmr/random.inp (Nilges et al., 1988a) and nmr/dgsa.inp (Nilges et al., 1988b) in X-PLOR 3.1; the starting point is a completely random set of Cartesian coordinates. Here we call the combined protocol MDSA-AM-RXYZ-1.0. We made only one modification to this protocol: increasing the weight factor of the NOE restraint term by a factor of three to improve convergence.

Calculation of dimer structures

We tested four protocols for generating dimer coiled-coil structures. The first two protocols, MDSA-SCC-RPP-1.0 and MDSA-SCC-RPP-1.1, are only slightly modified versions of the MDSA-SD-RPP-1.0 protocol previously described by Nilges (1993). These modifications are described in Table 2; for details of the MDSA-SD-RPP-1.0 protocol the reader is referred to the previous paper by Nilges (1993). These protocols assume no prior knowledge of the monomer structures. In contrast, the other two protocols we developed and tested, MDSA-SCC-WDMR-1.0 and MDSA-SCC-REFMR-1.0, exploit prior knowledge of the monomer structure to improve the convergence. The MDSA-SCC-WDMR-1.0 protocol is described in full detail below. The differences between this protocol and MDSA-SCC-REFMR-1.0 are described in Table 2.

The MDSA-SCC-WDMR-1.0 protocol

The protocol begins from a well-defined monomer structure by which we mean a structure with good geometrical energy and overall correct topology – in this case, α-helical. The monomers were calculated using MDSA-AM-RXYZ-1.0 with the unambiguous distance-restraint subsets. The monomer structure was initially orientated such that its geometric centre coincided with the origin, and its long axis lay along the x-axis. A second monomer was generated from this first one by rotating the coordinates 180° around the x-axis. The dimer was then refined in three stages: a high-temperature search stage, an annealing stage, and a minimisation stage.

To maintain the correct local structure of each monomer, the initial weights on the bond and angle terms, w_{bond} and w_{angle}, were set to 1.0; in addition, the hydrogen-bond and unambiguous distance restraints were maintained throughout using the more stringent square-well function for the NOE potential (effectively setting σ to ∞ in Eq. 7) with $w_{unambig}$ set initially to 0.02 and 0.16 for GCN4 and Jun, respectively. For the ambiguous restraints we used the soft NOE potential function with w_{ambig} set initially to 0.16 and 0.5 for GCN4 and Jun, respectively. Experimentally determined dihedral-angle ranges were restrained with an initial weight of $w_{cdih} = 0.05$. The symmetry of the dimer was enforced using the NCS and GSYM constraints, with randomised selection for GSYM. The initial value of the weights w_{NCS} and w_{GSYM} were 0.1 and 2.0, respectively, for GCN4 and 0.2 and 1.3, respectively, for Jun. This initial weighting on w_{NCS} is much higher than in the previous protocols; this means that the monomers are constrained to move cooperatively during the search stage. We used the 'HEPTAD' interaction term which restrains the geometric centres of each symmetry-related heptad to be within 10.4 Å using a square-well quadratic potential (Nilges and Brünger, 1991). Our justification for using this term is based on prior experimental evidence that GCN4 (O'Shea et al., 1991) and Jun form parallel coiled coils. This ensures that the two helices interact in a parallel manner, and hence is appropriate only for parallel coiled coils (a similar term could be used for antiparallel arrangements). The initial weight on the HEPTAD term was $w_{HEPTAD} = 0.1$.

For the search stage, the initial velocities were assigned randomly based on a Maxwellian distribution at 2000 K. All atoms were assigned a uniform heavy mass of 100 Da. To speed the calculation, nonbonded interactions were calculated only between C^{α} atoms, using an increased vdW radius (s = 1.2) and an initial vdW weighting factor of $w_{vdW} = 0.025$. For stability reasons, the planar dihedral weight was initially set to $w_{planar} = 0.1$. The following X-PLOR nonbonded parameters were used: CUTNB = 100 Å, TOLERANCE = 45 Å, and NBXMOD = +4. The system was coupled to a heat bath (2000 K) with a frictional coefficient of 10 ps^{-1}, and the trajectory of the system was then calculated for 100 ps using 5-fs time steps.

In the annealing stage, the repel potential was turned on for all atoms; the initial parameters were: CUTNB = 4.5 Å, TOLERANCE = 0.5 Å, NBXMOD = −3, s = 0.9, and $w_{vdW} = 0.00075$. The weight on the dihedral term was increased to $w_{CDIH} = 1.0$. The system was then cooled from 2000 to 100 K in 50-K decrements, with 1.3 ps of dynamics per decrement. After each decrement, the following parameters were multiplied by constant factors such that at the last decrement (when the temperature was 100 K), each parameter reached its final value indicated here: s =

TABLE 2
PROTOCOLS USED IN THIS PAPER

Protocol name[a]	Description
MDSA-AM-RXYZ-1.0	Protocols nmr/random.inp and nmr/dgsa.inp in X-PLOR 3.1 with the weight of the NOE term increased by a factor of 3
MDSA-SD-RPP-1.0	'Protocol 1' of Nilges (1993). Calculates a symmetric dimer starting from a chain with random ϕ-ψ angles
MDSA-SCC-RPP-1.0	= MDSA-SD-RPP-1.0 + the HEPTAD term; specific for symmetric coiled coils
MDSA-SCC-RPP-1.1	= MDSA-SCC-RPP-1.0 + randomised selection for GSYM
MDSA-SCC-WDMR-1.0	See text for a full description; starts with a well-defined monomer; local structure of the monomers is maintained during refinement, and the monomers move cooperatively
MDSA-SCC-REFMR-1.0	= MDSA-SCC-WDMR-1.0, except that w_{NCS} is set to 1.0 throughout the protocol, restricting the motion of the monomers to be highly cooperative. Designed to start with a refined monomer structure
MDSA-SMU-WDMU-1.0	Starts with a well-defined dimer or multimer structure produced by the above protocols; no search phase; re-annealing and minimisation with final weights and without the HEPTAD term
MDSA-SMU-WDMR-1.0	= MDSA-SCC-WDMR-1.0 without the HEPTAD term; suitable for any symmetric multimer where prior knowledge of the monomer structure is available
MDSA-SMU-REFMR-1.0	= MDSA-SCC-REFMR-1.0 without the HEPTAD term; suitable for any symmetric multimer where prior knowledge of the monomer structure is available

[a] We have named the protocols according to the following scheme: MDSA-aa-bbbb-nn, where MDSA stands for molecular dynamical simulated annealing; aa gives the molecular configuration: AM for an asymmetric molecule with no symmetry ambiguity; SMU for a symmetric multimer where at least some of the NOEs have symmetry ambiguity; SD for a symmetric dimer; SCC for symmetric coiled coils; bbbb gives the initial coordinates assumed: RXYZ, random x, y, and z coordinates; RPP, random ϕ and ψ angles; WDMR, well-defined monomer; WDMU, well-defined multimer; REFMR, refined monomer; REFM, refined molecule; nn indicates the version number of the protocol.

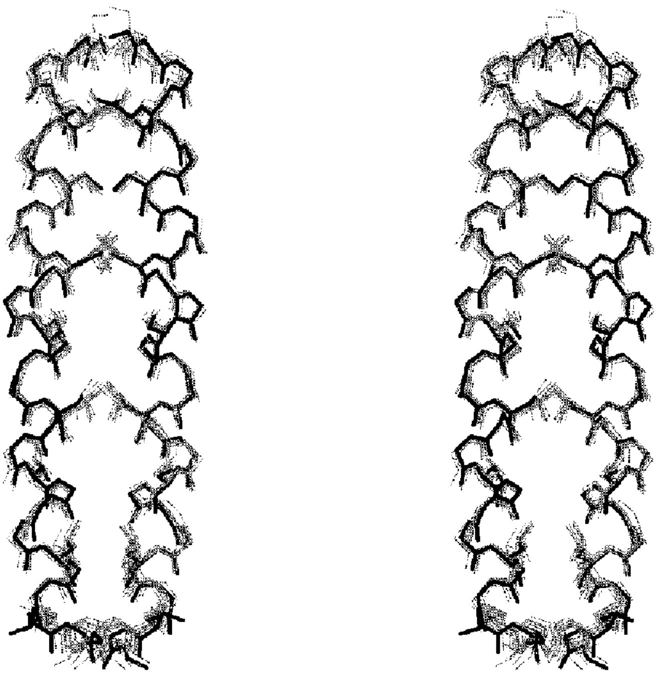

Fig. 1. Stereoview of the final 11 GCN4-ci structures showing all main-chain heavy atoms and all side-chain heavy atoms for a and d residues (grey lines). These structures were calculated using the MDSA-SCC-WDMR-1.0 protocol from distance data derived from the crystal structure of GCN4-LZ; the structures are superimposed onto the crystal structure (black lines). The N-terminus is at the top, and the C-terminus is at the bottom.

0.8, $w_{ambig} = 1.0$, $w_{CDIH} = 1.0$, $w_{GSYM} = 1.0$, $w_{HEPTAD} = 1.0$, $w_{NCS} = 2.0$, $w_{planar} = 1.0$, $w_{unambig} = 1.5$, and $w_{vdW} = 1.0$.

All nonbonded parameters and the energy weights were kept at these final values in the subsequent minimisation stage consisting of 500 cycles of Powell energy minimisation.

Refinement of dimer structures

Selected dimer structures were refined using the MDSA-SMU-WDMU-1.0 protocol. This consisted of a further annealing stage (from 2000 to 100 K as before) followed by 500 cycles of Powell energy minimisation. The HEPTAD term is turned off in this protocol. The weights of all energy terms were maintained at the final values throughout the protocol.

Intermolecular mean force potentials

Mean force potentials (MFP) were calculated using PROSA (Sippl, 1993). The total MFP for each residue was calculated in the presence of the other monomer. The intramolecular MFP was calculated in the absence of the other monomer. The intermolecular MFP is then the total MFP minus the intra-monomer MFP.

Results

Initial dimer calculations

We first tried to calculate the dimer structures directly without prior knowledge of the monomers using MDSA-SCC-RPP-1.0. However, the convergence rate was quite low: out of 50 structures calculated for each distance set, only seven for D^{Jun-m} and four for D^{Jun-s} converged to within 2.5 Å rmsd from Jun-m; three of the D^{GCN4-s} structures and only two for D^{GCN4-c} converged to within 2 Å rmsd from GCN4-c. In the majority of the structures which failed to converge, the two monomers had correctly separated from their initially coincident position, but had not undergone the 180° rotation necessary to satisfy the symmetry. In retrospect, we realised that for such a regular, extended molecule, the systematic GSYM selection produces this artefactual local minimum close to the trivial solution (i.e., where both monomers are completely coincident).

Thus, we tried simply switching to randomised selection for GSYM using MDSA-SCC-RPP-1.1. This improved the symmetry of the final structures, but the convergence rate was still quite low: seven out of 50 structures converged for D^{GCN4-c} (using the same convergence criteria as above) – this is not significantly higher than the convergence rate obtained with systematic selection ($p_s = 0.10$; see Appendix for an explanation of this statistic). We regarded this convergence rate as unacceptably low since these calculations were already slowed due to the high level of ambiguity in the data sets (since more distances have to be calculated). Therefore, we tried to improve the convergence by using prior information about the monomer structures.

Developing the protocol

Repeatedly applying the MDSA-AM-RXYZ-1.0 protocol using only the initially assigned intra-monomer distances, we generated 50 monomer structures for each distance set. The structures generated were completely α-helical, but the helix was twisted to varying degrees.

Using the model distance sets and starting from these monomer structures, we tested many variations on the initial dimer protocol hoping to obtain better convergence. Finally, we settled on MDSA-SCC-WDMR-1.0 which gave 30 and 28 out of 50 converged structures for $D^{Jun\text{-}m}$ and $D^{GCN4\text{-}c}$, respectively (same criteria as above) – clearly a highly significant improvement in convergence rate ($p_s < 10^{-6}$ and $p_s < 10^{-8}$, respectively) over that obtained with MDSA-SCC-RPP-1.0. Thus, at least for the model data sets, MDSA-SCC-WDMR-1.0 searches conformation space relatively efficiently and finds the correct solution.

Of the 50 $D^{GCN4\text{-}c}$ dimer structures, all in the top 50% (ranked in order of total energy) had the correct coiled-coil interface (a and d residues in the interface). We selected the 11 lowest energy structures as the final ensemble (denoted GCN4-ci, where i = 1,...,11); these had no NOE violations greater than 0.5 Å, and good covalent geometry (mean rmsd from ideal bond lengths, bond angles, and improper dihedral angles of 0.0025 ± 0.0002 Å, 0.38 ± 0.02°, and 0.39 ± 0.05°, respectively). The final ensemble superimposed closely onto GCN4-c with rmsd = 0.8 ± 0.1 Å and ρ = 0.060 ± 0.001 for main-chain atoms, rmsd = 1.9 ± 0.1 Å and ρ = 0.12 ± 0.01 for all atoms (Fig. 1).

In the case of Jun-m calculations, all structures in the top 50% also had the correct coiled-coil interface; however, in about half of these structures, the packing of one or two of the leucine residues was swapped. This indicates that the calculation had not yet converged to one structure. Further iteration of assignment and refinement stages would be necessary to obtain convergence. However, it was clear that the calculations were converging toward the correct structure since the two lowest-energy structures, and seven of the best 13 structures, had the correct packing. Thus, we selected these seven structures as the final $D^{Jun\text{-}m}$ ensemble, denoted Jun-mi (i = 1,...,7); these had no NOE violations greater than 0.5 Å, and good covalent geometry (mean rmsd from ideal bond lengths, bond angles, and improper dihedral angles of 0.0075 ± 0.0004 Å, 0.81 ± 0.03°, and 0.91 ± 0.05°, respectively). These structures superimposed closely onto Jun-m with an rmsd = 1.1 ± 0.1 Å and ρ = 0.060 ± 0.005 for main-chain atoms, and an rmsd = 2.0 ± 0.1 Å and ρ = 0.10 ± 0.01 for all atoms (Fig. 2).

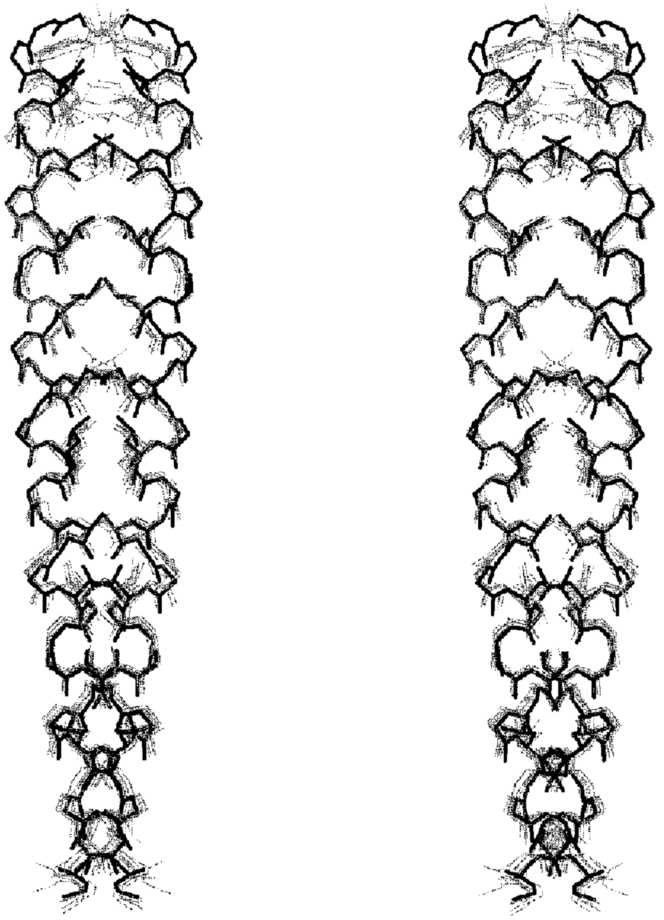

Fig. 2. Stereoview of the final seven Jun-mi structures showing all main-chain heavy atoms and all side-chain heavy atoms for a and d residues (grey lines). These structures were calculated using the MDSA-SCC-WDMR-1.0 protocol from distance data derived from a model structure of Jun-LZ; the structures are superimposed onto the model structure (black lines). The N-terminus is at the top, and the C-terminus is at the bottom.

Jun-s calculations

From the 50 monomer structures of Jun-s, we used MDSA-SCC-REFMR-1.0 to calculate dimer structures; again, the top 50% all had the correct coiled-coil interface. These were refined with MDSA-SMU-WDMU-1.0 and the 12 lowest-energy structures, denoted Jun-si, were selected as the final structures; these had no NOE violations greater than 0.5 Å, and good covalent geometry (mean rmsd from ideal bond lengths, bond angles, and improper dihedral angles of 0.0010±0.0001 Å, 0.18±0.01°, and 0.22±0.01°, respectively). These were all within 2.5 Å rmsd of Jun-m – a significant improvement over the results for MDSA-SCC-RPP-1.0 (p_s=0.03). The details of this structure will be presented elsewhere (Junius et al., 1996). Statistics on the assignments made from the final ensemble of structures are given in Table 1. In this case, unlike GCN4-s (below), some further inter- and co-monomer assignments could be made by additional rounds of calculation and assignment; however, the total number of assigned restraints per residue would be unlikely to increase much more.

GCN4-s calculations

The initial results for the $D^{GCN4\text{-}s}$ calculations were disappointing: only four out of 50 structures generated with MDSA-SCC-WDMR-1.0 converged to within 2.0 Å of GCN4-c – i.e., not significantly better than for MDSA-SCC-RPP-1.0. Moreover, 19 out of the 50 structures had equally low energies, but had clearly incorrect interfaces with many or all of the a and d residues completely exposed to the solvent. These problems arise mainly because $D^{GCN4\text{-}s}$ was originally derived by Saudek et al. (1991) with the intention of calculating only monomer structures: they excluded restraints which they suspected to be inter- or co-monomer. However, given that four of the dimer structures we calculated did have the correct interface, we suspected that this data set did contain some co-monomer NOEs. Thus, we proceeded with the calculation of a solution structure for GCN4 based only on the $D^{GCN4\text{-}s}$ data set using an iterative assignment strategy.

We excluded obviously bad structures from the ensemble of 50 by ranking in order of total energy, and also separately NOE energy; structures in the worse 20% of either list were removed. The remaining selected structures included a mixture of correct and incorrect interfaces: each ambiguous NOE was then checked against these structures. We assigned an NOE as intra- or inter-monomer only when every selected structure gave the same unambiguous assignment. Details of this iteration assignment method will be published elsewhere (Nilges, M., Macias, M., O'Donoghue, S.I. and Oschkinat, H., manuscript in preparation). Co-monomer NOE assignments were tested by making a NOE table of all possible co-monomer restraints, and testing which were not violated in all selected structures. Of the previously ambiguous assignments, 217 were unambiguously assigned as intra-monomer – 73 remained ambiguous. With this new set, we repeated the structure calculation, starting again from the 50 monomer structures. Again, excluding the

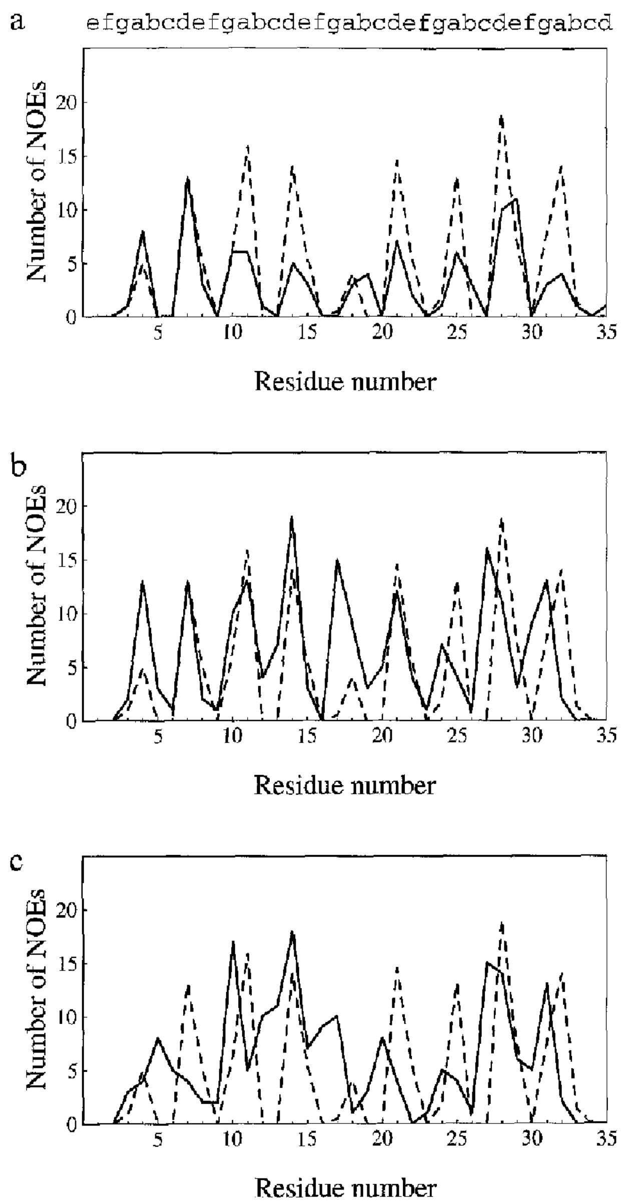

Fig. 3. (a) Correlation between the remaining ambiguous NOEs in $D^{GCN4\text{-}s}$ (solid line) with the number of inter- and co-monomer NOEs in $D^{GCN4\text{-}cd}$ (dotted line). The remaining ambiguous NOEs refer to the number after the second round of assignments in the GCN4-s calculation. The two distributions are very strongly correlated (Kendall's τ-test, $p \leq 10^{-9}$) suggesting that we can identify the interface residues in GCN4-s. (b) The correlation between the remaining ambiguous NOEs in $D^{GCN4\text{-}c\text{-}t2}$ (solid line) and the number of inter- and co-monomer NOEs in $D^{GCN4\text{-}cd}$ (dotted line). The $D^{GCN4\text{-}c\text{-}t2}$ data set was extracted from $D^{GCN4\text{-}c}$, which includes a small number of co-monomer constraints. The two distributions are also strongly correlated (p = 0.002). (c) The correlation between the remaining ambiguous NOEs in $D^{GCN4\text{-}c\text{-}t1}$ (solid line) versus the number of inter- and co-monomer NOEs in $D^{GCN4\text{-}cd}$ (dotted line). $D^{GCN4\text{-}c\text{-}t1}$ has only intra-monomer NOEs. Clearly there is a much poorer correlation (p = 0.37).

worst energy structures, we assigned an additional 10 NOEs as intra-monomer, leaving only 63 as ambiguous. The convergence of this second round towards GCN4-c was a little better than in the initial round, however still overall poor. Since we had so far been unable to assign a single inter- or co-monomer NOE, we judged that more rounds with this approach would be not likely to converge to a single dimer structure.

We compared the distribution of the 63 remaining ambiguous NOEs with distribution of inter- and co-monomer NOEs expected from $D^{GCN4\text{-}cd}$ (Fig. 3a) and we noticed a very strong correlation between them ($p = 10^{-9}$); this suggested that the set of residues with more than one remaining ambiguous NOE can be assumed to define the interface. That is, even though the convergence towards the correct dimer structure was very poor, the $D^{GCN4\text{-}s}$ data set still contains enough information to enable us to map which residues occur at the interface.

To check this interface-mapping idea, we set up two test calculations based on the $D^{GCN4\text{-}c}$ data set. For the first test, we constructed a data set, denoted $D^{GCN4\text{-}c\text{-}t1}$, which contained 356 NOEs (same size as $D^{GCN4\text{-}s}$) of which 22% could be initially assigned as unambiguous intra-monomer NOEs. The remaining ambiguous NOEs were randomly selected from the (initially unassignable) intra-monomer NOEs in $D^{GCN4\text{-}c}$ – i.e., no inter- or co-monomer NOEs. For the second test, we constructed another data set, denoted $D^{GCN4\text{-}c\text{-}t2}$, with the same size, and the same number of initially assignable intra-monomer NOEs, but with 8% of the NOEs randomly selected from the co-monomer NOEs in $D^{GCN4\text{-}c}$. The remaining ambiguous NOEs were selected from the initially unassignable intra-monomer NOEs as before. With these two data sets, we did two rounds of calculation and assignment, exactly as for $D^{GCN4\text{-}s}$. We then compared the distribution of remaining ambiguous NOEs to that of the inter- and co-monomer NOEs in $D^{GCN4\text{-}c}$ (Figs. 3b and 3c). Overall, the remaining ambiguous NOEs in $D^{GCN4\text{-}c\text{-}t1}$ was not significantly correlated with the distribution of inter- and co-monomer NOEs in $D^{GCN4\text{-}c}$ ($p = 0.37$). For a few residues, there does appear to be a correlation; this probably arises from the fact that interface residues have more intra-monomer NOEs than surface residues. In contrast, the distribution of remaining ambiguous NOEs in $D^{GCN4\text{-}c\text{-}t2}$ was highly correlated to the distribution of inter- and co-monomer NOEs in $D^{GCN4\text{-}c}$ ($p = 0.002$). Comparing this correlation with the correlation for $D^{GCN4\text{-}s}$ strongly suggests that $D^{GCN4\text{-}s}$ contains several co-monomer NOEs.

Hence, we felt justified in the assumption that we had identified the interface residues from $D^{GCN4\text{-}s}$ alone. The identified interface comprised all a and d residues and most b, e, and g residues. All c and f residues were excluded. This pattern is consistent with a classic coiled-coil packing throughout the entire molecule; involvement of b residues suggests the coiled-coil phase angle, θ (O'Donoghue et al., 1993), is greater than 26°, consistent with the GCN4-c structure.

Using this knowledge of the interface residues as a filter, we screened all the previous structures generated with $D^{GCN4\text{-}s}$ to select only those with the correct interface. We obtained nine structures from which we were able to assign eight co-monomer NOEs. With these co-monomer assignments, further structure calculations converged toward a single dimer structure. After seven more cycles of calculation and assignments, we had assigned 28 co-monomer NOEs and two inter-monomer NOEs; 11 restraints remained ambiguous. After a final refinement with the MDSA-SMU-WDMU-1.0, we selected the 19 lowest-energy structures, denoted GCN4-si ($i = 1,...,19$), as the final structures; these had no NOE violations greater than 0.5 Å, and good covalent geometry (mean rmsd from ideal bond lengths, bond angles, and improper dihedral angles of 0.0029 ± 0.0003 Å, 0.30 ± 0.01°, and 0.26 ± 0.01°, respectively). Statistics on the assignments made from the final ensemble of structures are given in Table 1.

The final assigned data set defines the backbone structure in residues 6 to 31 on both monomers (Fig. 4). In this range, comparing each GCN4-si structure with GCN4-c gives an rmsd = 1.6 ± 0.2 Å and $\rho = 0.13 \pm 0.02$ for main-chain atoms, and an rmsd = 2.8 ± 0.2 Å and $\rho = 0.22 \pm 0.01$ for all atoms; this indicates a relatively close agreement in overall main-chain structure to the crystal structure. Also in this range, the packing of interface residues (Fig. 4), and the inter-molecular MFP profile (Fig. 5) are very similar in the crystal and solution structures.

Discussion

In deriving LZ monomer structures, we have relied on the assumption that we can unambiguously assign many backbone–backbone NOEs as intra-monomer (O'Donoghue et al., 1993); these assignments are valid if the two monomers form helices separated by at least the known coiled-coil separation distance. No assumptions were made about which residues form the interface. Given the strong evidence we have in each case that these proteins form coiled coils, this should be a safe assumption. We justify our use of the HEPTAD constraint in deriving dimer structures with the same reasoning. If the data are derived from a parallel coiled coil, then this constraint acts only to increase convergence, but does not influence the final converged structure that is reached. The constraint drives both monomers to align in a parallel fashion - however once they are so aligned, the constraint is satisfied and no additional force is applied. Thus, the constraint is equally compatible with all coiled-coil pitch values, as well as non-coiled-coil geometries. The constraint would also allow a substantial slip between the monomers. The fact that in all cases the calculation converges to the correct coiled-coil pitch and phase indicates that the final structure is determined by the distance data. By contrast, in a previous modelling study of LZs (Nilges and Brünger, 1993), the force field included an attractive vdW term, and an electrostatic term; in that case, convergence towards the correct pitch and phase was driven by the force field.

The calculations using model distance sets suggest the protocol can determine LZ structures to an rmsd accuracy of around 0.9 Å for main-chain atoms, and 2.0 Å for all atoms. This accuracy seems poorer than we would first expect, since these are model calculations. To take a well-known example for comparison, the rmsd between the BPTI crystal structure (PDB code 4PTI) and the ensemble of BPTI solution structures (1PIT; Berndt et al., 1992) is 1.0±0.1 Å for main-chain atoms, and 1.7±0.1 Å for all atoms. This seems to suggest that the intrinsic accuracy of our protocol is the same magnitude as the expected accuracy of the NMR technique – ideally, the protocol should have a better accuracy. However, we need to consider that the LZs are very elongated molecules with unusually large radii of gyration. For such molecules, the rmsd measure over-estimates the differences. Recently, Maiorov and Crippen (1995) have proposed the size-independent ρ-factor as a more robust and unbiased mathematical measure of similarity for polypeptide chains. Using this measure, the average accuracy of the protocol (for Jun and GCN4) is 0.060±0.001 for main-chain atoms, and 0.11±0.01 for all atoms. Whereas, for the more globular BPTI, the ρ-factor difference between the crystal and solution structures is 0.10±0.01 for main-chain atoms, and 0.16±0.01 for all heavy atoms. That is, the intrinsic accuracy of the protocol with model data is better than the experimental accuracy of the NMR technique.

This level of accuracy clearly shows that the short-range distances used for the NMR structure determination technique can indeed accurately define a long-range property such as the overall pitch of the LZ coiled coil. The novel solution structure of the Jun LZ domain that we have generated using this protocol should have similar accuracy. These results, together with previous test calculations with three different structural classes of protein (Nilges, 1993,1995), further confirm the usefulness of dynamic assignment in calculating accurate structures, even when the data sets have a high degree of ambiguity.

The orientation of the two monomers in the GCN4-s solution structure calculated here depends on only 30 co-

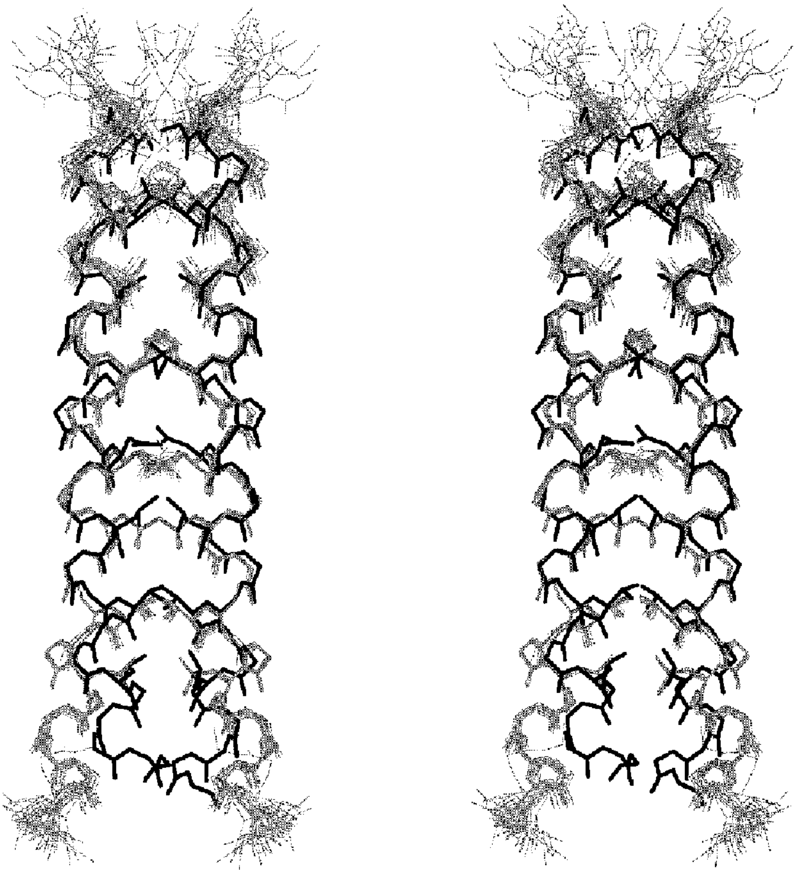

Fig. 4. Stereoview of the family of 19 GCN4 solution structures (grey lines) and the crystal structure (black lines) showing all main-chain heavy atoms and all side-chain heavy atoms for a and d residues. The N-terminus is at the top, and the C-terminus is at the bottom. The solution structures have been superimposed onto the crystal structure, matching the main-chain atoms in residues 6 to 31.

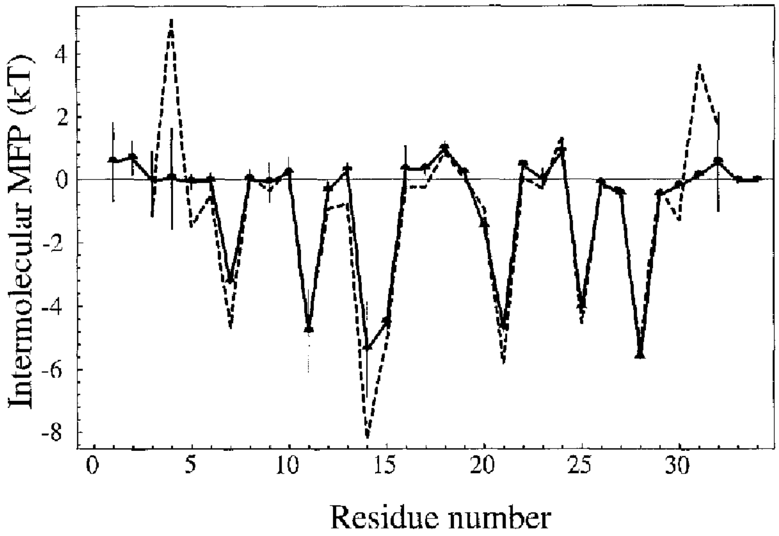

Fig. 5. Intermolecular mean force potentials calculated for the solution structures (solid line) and the crystal structure (dotted line) of GCN4. The error bars give the standard deviation of values over the family of 19 solution structures. For residues 6 to 31, the two potential profiles agree quite closely. The slightly positive value for the interface residue Asn[18] suggests that this residue is destabilising relative to the other interface residues.

and inter-monomer NOEs (less than one per residue); hence, we do not present this structure as a novel, atomic-resolution structure. Rather, we present these results firstly to illustrate the use of the calculation strategy: applied to the original NOESY spectra of Saudek et al. (1991) including the suspected inter- and co-monomer NOEs, the method would be likely to give a reasonably accurate solution structure of GCN4. Secondly, we draw some limited conclusions about the solution structure: the a and d residues interact along the whole length; the coiled-coil phase angle, θ, is probably greater than 26°; the packing of residues in the interface between residues 6 and 31 appears to be very similar to the crystal structure; the a-position asparagine residue appears to destabilise the dimer. These conclusions are in agreement with the GCN4 crystal structure.

As expected from the coincidence of the symmetry axis with the principal moment of the molecule, relatively few NOEs are purely inter-monomer for LZ domains, and there is a high proportion of co-monomer NOEs. This explains the difficulties we and others have had in generating dimer structures for LZs. Thus, even if isotopic labelling methods were used, it would still be necessary to use dynamic assignment and co-monomer restraints. The novel methods presented here (co-monomer restraints and interface mapping) may be useful in solving other multimers with difficult symmetries.

Since the heptad constraint was used in both the MDSA-SCC-RPP-1.0 and MDSA-SCC-WDMR-1.0 protocols, the dramatic difference in their performance is probably due to starting with the monomer structure. This suggests that the MDSA-SMU-WDMR-1.0 protocol (i.e., MDSA-SCC-WDMR-1.0 without the HEPTAD term*) may be the more useful in cases where data on the monomer structure are available. Where no prior knowledge of the monomer is available, the previous protocol MDSA-SD-RPP-1.0 is still the method of choice. Our experience also cautions us that protocols developed using particular protein structure classes may not work for other protein classes. Hence, the protocols should be regarded only as prototypes for the general case of symmetric multimers.

Conclusions

We have shown that the NMR structure determination method can indeed produce correct structures even for the difficult case of the extended, symmetric LZ domain. This is demonstrated with the first complete solution structure of an LZ domain – the homodimer of Jun-s. The novel calculation methods presented here may be useful for studying other symmetric multimers where many NOEs are co-monomer, or where prior knowledge of monomer structures is available.

Acknowledgements

We would like to thank Dr. Keith Junius for providing NMR data on Jun, and for many helpful discussions and advice. We would also like to thank Dr. Annalisa Pastore for making available the NMR data on GCN4. This work was supported by a C.J. Martin postdoctoral fellowship to S.I.O'D. from the National Health and Medical Research Council of Australia.

*In cases of other multimers, it may be useful to include a more generalised form of the HEPTAD restraint which would prevent monomers from drifting apart, hence improving convergence.

References

Arrowsmith, C.H., Pachter, R., Altman, R.B., Iyer, S.B. and Jardetzky, O. (1991) *Biochemistry*, **29**, 6332–6341.

Atkinson, R.A., Saudek, V., Huggins, J.P. and Pelton, J.T. (1991) *Biochemistry*, **30**, 9387–9395.

Berndt, K.D., Güntert, P., Orbons, L.P.M. and Wüthrich, K. (1992) *J. Mol. Biol.*, **227**, 757–775.

Brünger, A.T., Clore, G.M., Gronenborn, A.M. and Karplus, M. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 3801–3805.

Brünger, A.T. (1992) *X-PLOR v. 3.1, User Manual*, Yale University, New Haven, CT.

Clore, G.M., Omichinski, J.G., Sakaguchi, K., Zambrano, N., Appella, E. and Gronenborn, A. (1995) *Science*, **267**, 1515–1516.

Crick, F.H.C. (1953) *Acta Crystallogr.*, **A6**, 685–689.

Folkers, P.J.M., Folmer, R.H.A., Konings, R.N.H. and Hilbers, C.W. (1993) *J. Am. Chem. Soc.*, **202**, 3798–3799.

Folmer, R.H.A., Hilbers, C.W., Konings, R.N.H. and Hallenga, K. (1995) *J. Biomol. NMR*, **5**, 427–432.

Havel, T.F. and Wüthrich, K. (1984) *Bull. Math. Biol.*, **46**, 673–698.

Junius, F.K., Weiss, A.S. and King, G.F. (1993) *Eur. J. Biochem.*, **214**, 415–424.

Junius, F.K., O'Donoghue, S.I., Nilges, M., Weiss, A.S. and King, G.F. (1996) *J. Biol. Chem.*, **271**, 13663–13667.

Lee, W.T., Harvey, T.S., Yin, Y., Morin, P., Litchfeld, D. and Arrowsmith, C.H. (1995) *Nature Struct. Biol.*, **1**, 877.

Levy, R.M., Bassolino, D.A., Kitchen, D.B. and Pardi, A. (1989) *Biochemistry*, **28**, 9361–9372.

Maiorov, V.N. and Crippen, G.M. (1995) *Proteins*, **22**, 273–283.

Nilges, M., Clore, G.M. and Gronenborn, A.M. (1988a) *FEBS Lett.*, **239**, 129–136.

Nilges, M., Clore, G.M. and Gronenborn, A.M. (1988b) *FEBS Lett.*, **229**, 317–324.

Nilges, M., Gronenborn, A.M., Brünger, A.T. and Clore, G.M. (1988c) *Protein Eng.*, **2**, 27–38.

Nilges, M. and Brünger, A.T. (1991) *Protein Eng.*, **4**, 649–659.

Nilges, M. (1993) *Proteins*, **17**, 297–309.

Nilges, M. and Brünger, A.T. (1993) *Proteins*, **15**, 133–146.

Nilges, M. (1995) *J. Mol. Biol.*, **245**, 645–660.

Oas, T.G., McIntosh, L.P., O'Shea, E.K., Dahlquist, F.W. and Kim, P.S. (1990) *Biochemistry*, **29**, 2891–2894.

O'Donoghue, S.I., Junius, F.K. and King, G.F. (1993) *Protein Eng.*, **6**, 557–564.

O'Shea, E.K., Klemm, J.D., Kim, P.S. and Alber, T. (1991) *Science*, **254**, 539–543.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1986) *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Cambridge, U.K.

Saudek, V., Pastore, A., Castiglione Morelli, M.A., Frank, R., Gausepohl, H., Gibson, T., Weih, F. and Roesch, P. (1990) *Protein Eng.*, **4**, 3–10.

Saudek, V., Pastore, A., Castiglione Morelli, M.A., Frank, R. and Gibson, T. (1991) *Protein Eng.*, **4**, 519–529.

Sippl, M. (1993) *Proteins Struct. Funct. Genet.*, **17**, 355–362.

Wüthrich, K., Billeter, M. and Braun, W. (1983) *J. Mol. Biol.*, **169**, 949–961.

Appendix

Binomial statistics for comparing the convergence of algorithms

Here we address the question of how to compare the convergences of two algorithms and decide if the difference is statistically significant. We assume in this discussion that the convergence of an algorithm is represented in a binary way, i.e. for each trial calculation of the algorithm, we consider that the result is either successful (converged) or not successful. For NMR structure determination this is usually the case, since we normally have some definite acceptance criteria that structures either pass or fail.

Assume we have two algorithms for which we have calculated n_1 and n_2 structures, of which r_1 and r_2, respectively, have converged; the estimated convergence rates for the two algorithms are then $p_1 = r_1/n_1$ and $p_2 = r_2/n_2$. Without loss of generality, we assume that $p_1 > p_2$ (when $p_1 = p_2$ there is clearly no significant difference). Is the difference $p_1 - p_2$ significant? That is, do the data show a significant difference in the convergence rate of the two algorithms?

We take as a null hypothesis that both algorithms have the same convergence rate; the most likely estimate of this rate is:

$$p = \frac{r_1 + r_2}{n_1 + n_2} \tag{A-1}$$

According to the null hypothesis, the observed deviance of p_1 and p_2 from p occurs purely by chance. Thus, the significance level is the probability of the observed deviance or worse by chance alone. This is calculated from the probability of obtaining $\geq r_1$ successes or $\leq r'_1 = \text{round}(2n_1 p - r_1)$ successes with n_1 trials, and $\geq r_2$ successes or $\leq r'_2 = \text{round}(2n_2 p - r_2)$ successes with n_2 trials, where 'round' indicates rounding to the nearest integer number. Thus, the significance level is calculated using:

$$\begin{aligned} p_s = & \left[\sum_{r=0}^{r'_1} \binom{n_1}{r} p^r (1-p)^{n_1-r} + \sum_{r=r_1}^{n_1} \binom{n_1}{r} p^r (1-p)^{n_1-r} \right] \\ & \times \left[\sum_{r=0}^{r_2} \binom{n_2}{r} p^r (1-p)^{n_2-r} + \sum_{r=r'_2}^{n_2} \binom{n_2}{r} p^r (1-p)^{n_2-r} \right] \end{aligned} \tag{A-2}$$

where

$$\binom{n}{r} = \frac{n!}{(n-r)!\, r!} \tag{A-3}$$

When $p_s \leq 0.05$, we reject the null hypothesis in favour of the alternative hypothesis that the two algorithms have different convergence rates.