Journal of Biomolecular NMR, 7 (1996) 48–58
ESCOM

J-Bio NMR 324

Validation of the use of intermolecular NOE constraints for obtaining docked structures of protein–ligand complexes

Michael J. Gradwell and James Feeney*

Laboratory of Molecular Structure, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, U.K.

Received 22 August 1995
Accepted 28 November 1995

Keywords: NOE; Docking; Accuracy; Precision; R^{-6} averaging; Simulated annealing; Rigid-body dynamics

Summary

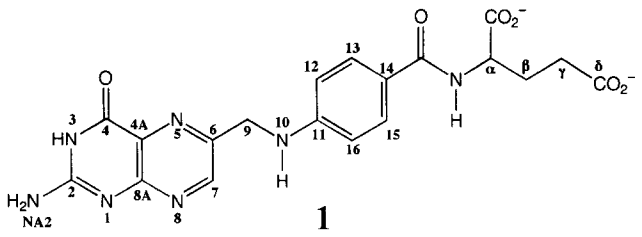
The use of intermolecular NOEs for docking a small ligand molecule into its target protein has been investigated with the aim of determining the effectiveness and methodology of this type of NOE docking calculation. A high-resolution X-ray structure of a protein–ligand complex has been used to simulate loose distance constraints of varying degrees of quality, typical of those estimated from experimental NOE intensities. These simulated data were used to examine the effect of the number, distribution and representation of the experimental constraints on the precision and accuracy of the calculated structures. A standard simulated annealing protocol was used, as well as a more novel method based on rigid-body dynamics. The results showed some analogies with those from similar studies on complete protein NMR structure determinations, but it was found that more constraints per torsion angle are required to define docked structures of similar quality. The effectiveness of different NOE-constraint averaging methods was explored and the benefits of using 'R^{-6} averaging' rather than 'centre averaging' with small sets of NOE constraints were shown. The starting protein structure used in docking calculations was obtained from previous X-ray or NMR structure studies on a related complex. The effects on the calculated conformations of introducing structural differences into the binding site of the initial protein structure were also considered.

Introduction

In functional and inhibition studies of enzymes, it is necessary to obtain structural information for the proteins in their complexes with a range of substrates and inhibitors. It is possible to determine 3D models of such complexes, with good accuracy, using either X-ray crystallography or NMR spectroscopy. For large proteins the NMR-based determination of a complete structure for each complex in solution can be time-consuming and, in some cases, may not be possible. However, the conformation and environment of the bound ligand can sometimes be deduced by an alternative method, where the structure of the complex is determined by first measuring a set of intermolecular ligand–protein NOEs and then using these as distance constraints to dock the ligand molecule into the binding site of a protein structure previously determined for a related complex (Bennion et al., 1992; Fesik et al., 1992; Weber et al., 1992,1993,1994; Lian et al., 1994; Martorell et al., 1994; Byeon et al., 1995). Such a method can be justified where the overall structure of the protein in the two complexes can be shown to be very similar, which is often the case. It is particularly useful for large proteins, where it is sometimes difficult to obtain sufficient NMR constraints to allow a complete structure determination.

The simplest methods for obtaining docked structures of this type have used interactive molecular graphics to place the ligand manually within the protein binding site in an orientation estimated to satisfy the experimental constraints. These coordinates are then subjected to restrained energy minimization (Weber et al., 1992,1993, 1994; Martorell et al., 1994; Byeon et al., 1995), or sometimes restrained molecular dynamics (RMD) (Fesik et al., 1992), to yield the final docked structure. Semi-automatic methods have also been reported, where RMD has been

*To whom correspondence should be addressed.



Scheme 1. The structure of folic acid.

used to dock a ligand that was positioned outside the binding site at the start of the calculation (Bennion et al., 1992). More detailed analyses have suggested that the use of a Monte Carlo method (Metropolis et al., 1953), such as simulated annealing (Kirkpatrick et al., 1983), is a highly efficient and readily implementable way of automatic docking using intermolecular distance constraints (Yue, 1990; Lian et al., 1994; Martorell et al., 1994). Such methods are robust and eliminate user bias. They allow the ligand and the protein to be at any initial location and orientation with respect to each other. They also allow a search of the conformational space available to the ligand as a function of the restraints, albeit to a limited extent, since it is necessary for at least some of the protein coordinates, such as its backbone atoms, to be kept fixed if only a few or none of the protein–protein distance constraints are available. Some other, more unusual methodologies for distance-constrained molecular docking have also been suggested, such as an ellipsoid algorithm by Billeter and co-workers (1987). Meadows and Hajduk (1995) have investigated the use of a genetic algorithm approach for small-ligand NOE docking.

Investigations of the accuracy of protein models calculated a priori from NMR data, and their dependence on the quantity and quality of the constraints used, have been reported (Clore et al., 1986,1993; Liu et al., 1992; Zhao and Jardetzky, 1994). This paper describes an investigation of this type for ligand–protein docking using only intermolecular constraints. It is aimed at assessing the fidelity of the results of such calculations and their dependencies on the number, quality and spatial distribution of the distance constraints. The crystal structure of the human form of dihydrofolate reductase (DHFR) in its binary complex with folic acid has been used to simulate sets of suitable distance constraints (Oefner et al., 1988). The molecular mass of this protein is 22 kDa and it is therefore relatively large for NMR studies, making it a typical candidate for ligand docking. The folate ligand consists of two rigid, cyclic systems joined by a flexible methylene linkage, terminated with a glutamic acid moiety (structure **1**, see Scheme 1). It therefore comprises several groups with differing degrees of conformational freedom. Structures have been calculated in XPLOR 3.1 (Brünger, 1992) using a standard simulated annealing protocol for NMR structure determinations (Nilges et al., 1988; Brünger, 1992). A novel simulated annealing method based on using a rigid-body dynamics approach (Yue, 1990) has also been tested. The use of inverse sixth power averaging (Brünger et al., 1986; Levy et al., 1989; Constantine et al., 1992) for considering the distances involving exchanging protons instead of the more conventional centre averaging (Brünger, 1992) or 'pseudoatom' approach has also been explored, since this method has been reported to give better defined structures in docking calculations (Lian et al., 1994).

Methods

The structure calculations were carried out on either a Sun Sparcstation 10 model 41 or a Silicon Graphics Iris Indigo R4000 XZ using XPLOR v. 3.1 (Brünger, 1992). Visualisation of structures was performed using the InsightII program within the Biosym package (Biosym Technologies, San Diego, CA), running on either Silicon Graphics Iris Indigo R4000 XZ or R4000 Elan workstations. For the definition of protein energy and topology within XPLOR, the standard tables 'parallhdg.pro' and 'topallhdg.pro' (Brünger, 1992) were used. These are designed for NMR structure determinations and utilise large force constants on the geometric parameters of the protein. Analogous user-defined data sets were written for the ligand, folate. The heavy atom coordinates of the crystal structure of human DHFR bound to folate (Oefner et al., 1988) were obtained from the Brookhaven Protein Data Bank. Hydrogen atoms were added to this structure using the Hbuild subroutine (Brünger and Karplus, 1988) within XPLOR. The geometry of the structure was subsequently regularized by performing 500 cycles of unrestrained Powell energy minimization (Powell, 1977). These coordinates were used for the generation of starting structures and for the calculations of the simulated distance constraints. A 'library' of initial positions for the folate ligand was generated by taking the original coordinates and subjecting the ligand to translation along a random vector in 3D space. The ligand was then assigned an arbitrary conformation by subjecting it to several picoseconds of free dynamics at a very high temperature (4000 K), with a simple nonbonded hard-sphere potential for atom–atom interactions (as in the simulated annealing calculations). Finally, it was rotated by a set of three randomized Eulerian angles. This allowed a whole range of initial positions and conformations for the ligand to be sampled.

For the display of structure sets and calculations of rms deviations, groups of calculated conformations were superposed on the original coordinates, using the backbone atoms of the protein. The graph plots were created using the GNUPLOT 3.5 package (copyright T. Williams and C. Kelley, Dartmouth).

Determination of distance restraints

Simulated NOE data sets were determined, using a

program written in ANSI C, by calculating all distances between protein and ligand protons within a cutoff of 5.0 Å. Each of these distances was then allocated to the appropriate range, corresponding to either 0–2.7 Å, 0–3.5 Å or 0–5.0 Å. This simulated the typical classification of NOEs as 'strong', 'medium' or 'weak'. Distance constraints to protons in methylene groups and to the two methyl substituents of valines and leucines in proteins are usually averaged in structure calculations if the stereospecific assignments are unknown ('pseudoatoms'; Wüthrich et al., 1983). Similar averaging is also often used for protons in NH_2 groups and in methyl groups such as those in alanines and isoleucines, and for δ and ε pairs of protons in tyrosines and phenylalanines, since their rates of conformational exchange tend to render them equivalent on the NMR time scale. The data set was adjusted to allow for this averaging: with the exception of β methylenes, the shortest simulated distance for each of the above types of groups was used, and the appropriate distance correction was added to the upper limit (Wüthrich et al., 1983) to allow for averaging of the constraint to the geometric centre ('centre averaging'; Brünger, 1992) of each of these groups ('pseudoatom correction'). In order to assess the dependence of the results on data quality, protons on the ligand were considered in three different ways. Firstly, a set of simulated restraints was determined with each hydrogen on folate discretely assigned (type 1 data set). In a second data set, the two groups of methylene protons in the glutamate moiety, as well as the pairs of hydrogens on C9 and the NH_2 substituent were centre averaged (type 2 data set). In the third set of constraints, centre averaging between the two pairs of aromatic protons (H13 and H15, and H12 and H16) attached to the benzoyl ring (type 3 data set) was also included. This simulated the situation where the rate of ring flipping of the benzoyl ring would be such that the signals for H13 and H15, for example, coalesce, as discussed above. Figure 1 summarizes the number and distribution of constraints for each data type.

Structure calculation methodologies

A standard simulated annealing protocol provided with XPLOR v. 3.1 ('sa.inp', Nilges et al., 1988; Brünger, 1992) was employed to calculate docked structures using the three sets of constraints for the parent data sets. This method utilises a soft, square-well potential for the NOE energy function, with initially a gentle slope on the asymptote of this function, and a very small force constant for the nonbonded interactions (a hard-sphere potential). After 5 ps of dynamics at 1000 K, the curvature of the asymptote is increased, and a further 2.5 ps of dynamics are calculated. The system is slowly, linearly,

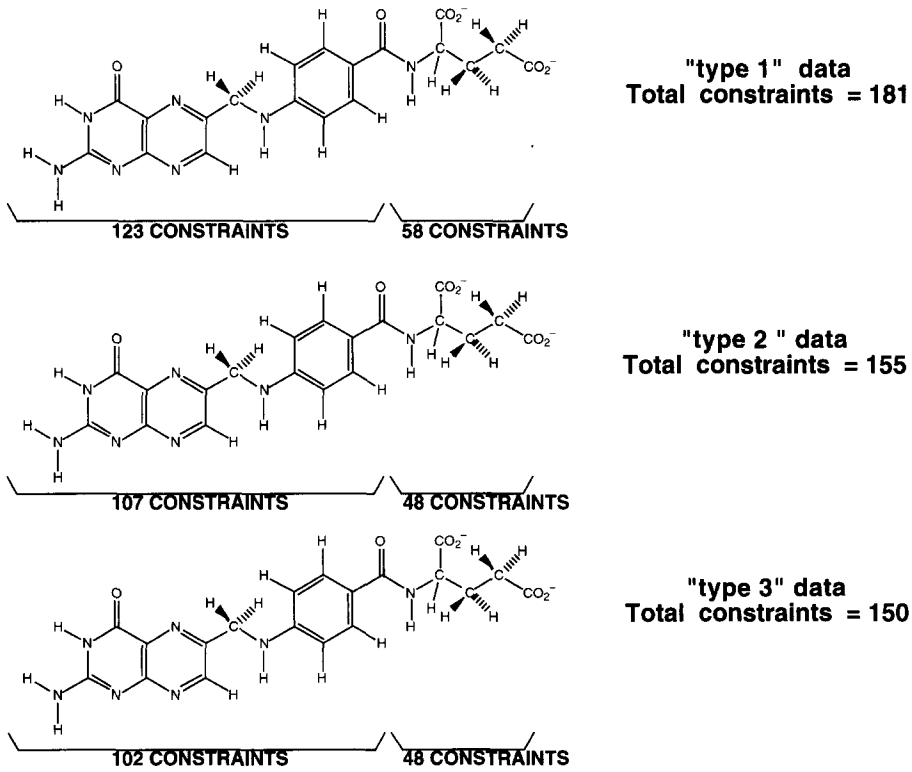


Fig. 1. Summary of the distribution of all possible simulated NOE constraints among the two ring systems and the glutamate moiety. The constraints were calculated from the X-ray structure using a 5.0 Å cutoff. Three data types are defined, based on using different methods of averaging within certain groups of ligand atoms. 'Type 1' data have constraints obtained with no averaging of ligand protons. 'Type 2' data have constraints obtained with averaging of pairs of ligand methylene and NH_2 protons. 'Type 3' data have constraints obtained with averaging of the methylene protons, NH_2 protons and pairs of aromatic ring protons.

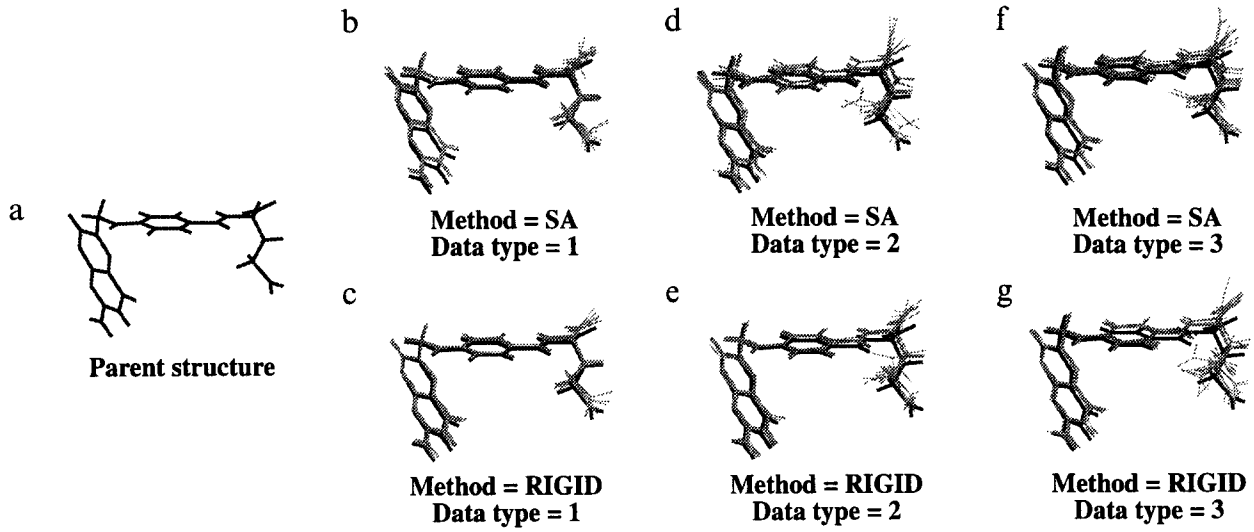


Fig. 2. Docked ligand structures calculated using all possible simulated distance constraints. For clarity, the protein atoms are omitted. (a) 'Parent' structure of folate in its complex with DHFR as determined by X-ray crystallography. (b)–(g) In each of these diagrams, 10 calculated structures are shown superimposed (using protein backbone atoms) with the original crystal structure (dark line). The calculated conformations are shown in lighter lines. The data type used, based on averaging of the ligand protons, is indicated for each structure set (see Fig. 1). The two different docking protocols are denoted by 'SA' or 'RIGID' (see the text for details).

cooled to 100 K, while the force constant for the nonbonded interactions is scaled up (with simultaneous reduction in the hard-sphere radii) exponentially. A total of 4 ps of dynamics are performed for the duration of the cooling. In the simplest version of this protocol, the protein atoms were kept fixed during the various dynamics stages. Finally, 200 cycles of restrained Powell energy minimization (Powell, 1977) were calculated with all atoms free to move. This method was designed to converge coordinates to their global energy minimum from any arbitrary initial configuration. It was thus found to be sufficiently robust to dock the ligand from any initial position; the initially small force constant for the nonbonded interaction allowed the folate ligand to pass through the atoms of the protein. This calculation strategy is described here as the 'SA' method.

A more specialised protocol was also designed and tested. This method utilises rigid-body dynamics (Brünger, 1992). It has been demonstrated (Yue, 1990) that this is a highly efficient way of obtaining two docked species in their approximately correct, relative orientations with respect to each other, with some further dynamics possibly being necessary to obtain the ligand in its final correct conformation. Using a target temperature of 1000 K, 2 ps of restrained rigid-body dynamics were calculated, with the ligand and the protein being considered as two rigid groups. This was always found to be sufficient to move folate into the binding site and allow equilibration. It was followed by a further period of 'normal' molecular dynamics at 1000 K for 2.5 ps. In the initial studies, the protein was held fixed and the ligand was able to move freely into the correct global minimum energy conformation. During these two periods, the hard-sphere nonbonded potential was scaled down, as above. This method allows the harsher square-well penalty function to be used for the distance constraints throughout the calculation. The next stage of the protocol consisted of standard, slow-cooling dynamics ('refine.inp'; Brünger, 1992); the target temperature was reduced from 1000 K to 100 K in 50 K steps, with scaling of the nonbonded potential as described above. As in the SA method, a total of 4 ps of dynamics were performed for the cooling. The final stage involved 200 cycles of restrained energy minimization, with all atoms able to move freely. This protocol is described here as the 'RIGID' method. Rigid-body dynamics proved to be better than rigid-body minimization for locating the ligand into its binding site when only sparse constraints were available.

During the calculations, a relatively high force constant (100 kcal mol^{-1}) for the NOE constraint energy function was used. The time step used in the molecular dynamics was 2.5 fs. Bond lengths were constrained using the SHAKE algorithm (Rykaert et al., 1977).

Results and Discussion

Ten structures with good energies, satisfying the restraints to a strict cutoff (0.1 Å), were calculated for each full data set from types 1, 2 and 3 (181, 155 and 150 constraints, respectively) using the SA (Figs. 2b, d and f) and RIGID (Figs. 2c, e and g) methods. When all the possible intermolecular NOE restraints are used (type 1), it can be seen that both protocols reproduce the correct ligand configuration and conformation, with approximately the same good levels of precision and accuracy. However, in addition to being more robust, the novel

RIGID protocol produced structures more quickly and had a better rate of convergence; typically, the calculation of a structure took circa twice as long using the SA rather than the RIGID protocol with the same data set. The yield of converged structures was approximately 10% more with the RIGID procedure, although this depended on the data set used. With large quantities of constraints, both methods achieved essentially 100% yields of converged structures, but with smaller data sets (< 50 constraints) the yields were 60–70%. However, the RIGID protocol was able to cope better with very small sets of restraints, particularly for cases where R^{-6} averaging was used (see below). Both methods were used to assess the dependence of the results on the quantity of available data, given an equivalent spatial distribution of constraints.

The effects of varying the number of constraints

A subset of 50 restraints was randomly chosen from the parent, type 3 data set. It was ensured that this subset had approximately the same distribution of distance constraints across the ligand as that of the full set (34 of the 50 constraints involved the nonglutamate part of the ligand) (Fig. 1). This file was used to create an analogous table of restraints for the other two data types (types 1 and 2) by changing the pseudoatoms to their correct discretely assigned protons where appropriate, and removing the corresponding distance corrections from the upper limit. As before, 10 structures for each set of constraints were calculated using both the SA and RIGID methods. The number of structures calculated was relatively small. However, it was found that only minor statistical improvements were obtained by examining larger families of structures (20–25). Consequently, it was concluded that families of 10 structures were sufficient for demonstrating gross overall trends. The resulting six sets of conformations are shown superimposed on the original crystal structure in Figs. 3b–g. It is clear that similar guidelines as those described previously can be derived for assessing the validity of full NMR structure determinations (Clore et al., 1986,1993; Liu et al., 1992; Zhao and Jardetzky, 1994). Clore and co-workers (1993) described how the overall quality of calculated protein conformations can be improved if stereospecific assignments are available. A similar trend is observed in our results (see Fig. 3). The determination of stereospecific assignments for methylene groups in a small molecule bound to a protein is nontrivial. Although obtaining a complete type 1 data set as defined in this study is unrealistic, it does act as a useful control. Specific assignments of aromatic ring protons are also important. By recording NMR spectra on samples at low temperatures, it is sometimes possible to decrease the rate of aromatic ring flipping in a ligand sufficiently to allow protons from the pairs of protons on either side of the ring to give rise to separate resonances. In this case, one can sometimes assign separate constraints to each of the protons in the pair (assigned after an initial round of calculations). For the more general case, the ring is flip-

TABLE 1
MEANS AND STANDARD DEVIATIONS FOR ENSEMBLES OF DOCKED LIGAND STRUCTURES CALCULATED WITH DIFFERENT NUMBERS OF NOE CONSTRAINTS (CENTRE AVERAGED) OBTAINED FROM THE PARENT STRUCTURE

Data type	Number of constraints[a]		Mean rms deviation (Å)	
	complete ligand	excluding glutamate moiety	subset heavy atoms[b] (SA)[c]	subset heavy atoms[b] (RIGID)[c]
1	181 (All)	123 (All)	0.24±0.07	0.14±0.04
1	70	50	0.22±0.05	0.20±0.06
1	60	38	0.29±0.07	0.28±0.05
1	50	34	0.24±0.07	0.24±0.07
1	40	28	0.44±0.06	0.40±0.08
1	30	22	0.88±0.40	0.82±0.34
2	155 (All)	107 (All)	0.32±0.09	0.21±0.04
2	70	50	0.29±0.07	0.28±0.10
2	60	38	0.33±0.08	0.26±0.07
2	50	34	0.35±0.08	0.38±0.12
2	40	28	0.43±0.09	0.69±0.42
2	30	22	0.65±0.28	0.70±0.41
3	150 (All)	102 (All)	0.35±0.11	0.26±0.06
3	70	50	0.34±0.08	0.37±0.09
3	60	38	0.43±0.10	0.44±0.10
3	50	34	0.54±0.42	0.64±0.47
3	40	28	0.77±0.23	0.85±0.39
3	30	22	0.92±0.42	1.14±0.31

[a] All data subsets had approximately the same distribution of NOEs across the folate molecule as the parent sets.
[b] Atoms of the flexible glutamate moiety were not included in the calculation of the rms deviation.
[c] This refers to the docking protocol used (see text for details).

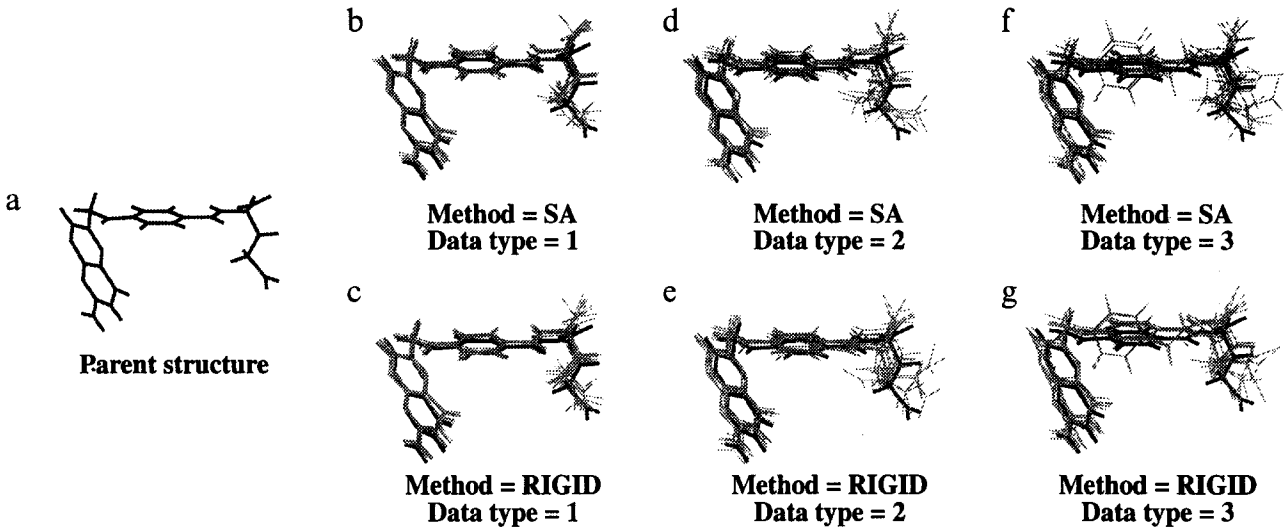


Fig. 3. Docked ligand structures calculated using the 50-constraint subsets of the parent data, with the same distribution across the ligand (34 of the 50 constraints were for the fragment without the glutamate moiety). The details of a–g are as for Fig. 2.

ping and it is necessary to use an averaging method. Given the larger distance between a pair of aromatic ring protons (for example the H13 and H15 protons) compared with that in a methylene group, the need for R^{-6} averaging rather than centre averaging is much more important in this case. This is well illustrated by the results shown in Fig. 3, where the type 3 data set fails to define a unique ligand conformation around the two rings, whereas by using data types 1 and 2 the torsional angles involved are reasonably well defined. However, in the case of the structures calculated using the type 2 constraints, the end of the glutamate chain is seen to occupy a whole range of conformations. To a lesser extent, this is observed even if the full data set is used for both types 2 and 3 (Fig. 2). The reason for this is that the glutamate moiety has less intermolecular distance restraints per torsion angle than the other parts of the ligand, even for the type 1 data set (Fig. 1). This part of the ligand was not included in subsequent quantitative studies.

Data sets with 30, 40, 60 and 70 constraints (corresponding to 22, 28, 38 and 50 constraints for the nonglutamate part of the ligand) were also used to calculate sets of structures analogous to those described above. Using the criterion for superposition described in the Methods section, the mean rms deviations between groups of ligand heavy atoms in the calculated structures and in the original model structure were calculated. These averages and their standard deviations give a measure of accuracy and precision, respectively. The values obtained for each family of structures are shown in Table 1. Overall, the SA and RIGID methods give generally similar results, indicating that either protocol is suitable for docking, although there are some minor differences. For example, the RIGID protocol gives structures with a slightly better accuracy and precision when large data sets are used. In contrast, results obtained with some of the data sets with very sparse constraints are a little better when the SA protocol is used. The mean rms deviations for the heavy atoms in the nonglutamate part of the ligand from their positions in the parent structure are correlated with the number of constraints. Plots of the rms deviations for this ligand heavy atom subset versus the total number of constraints are shown in Fig. 4. The dependence is seen to be approximately monotonic and is analogous to functions derived for complete structure determinations (Clore et al., 1986,1993; Liu et al., 1992; Zhao and Jardetzky, 1994). There are, however, some interesting differences. Most importantly, in addition to its overall conformation, the orientation of the ligand in its binding site in the protein is determined only by the intermolecular constraints, as the ligand has no covalent attachments to the protein structure. As a result of the increased freedom of the ligand, the dependence of the overall accuracy and precision on the number of constraints is much stronger, and their degradation with decreasing numbers of restraints is therefore more severe.

As the number of constraints is reduced, the calculated structures are seen to cluster around several discrete conformations, one of which corresponds to the true structure (see Figs. 3f and g), before degrading to a whole range of possible conformers. This is particularly apparent for the conformations of the two ring systems (described by three torsion angles), where no discrete solutions are identifiable when less than about 20–30 restraints are used for this part of the ligand, with centre averaging of constraints. However, if the structures are inspected closely, it can be seen that this finding also applies to the more flexible glutamate chain. This indicates the danger of overinterpretation, where the existence of true multiple conformations in the bound ligand may

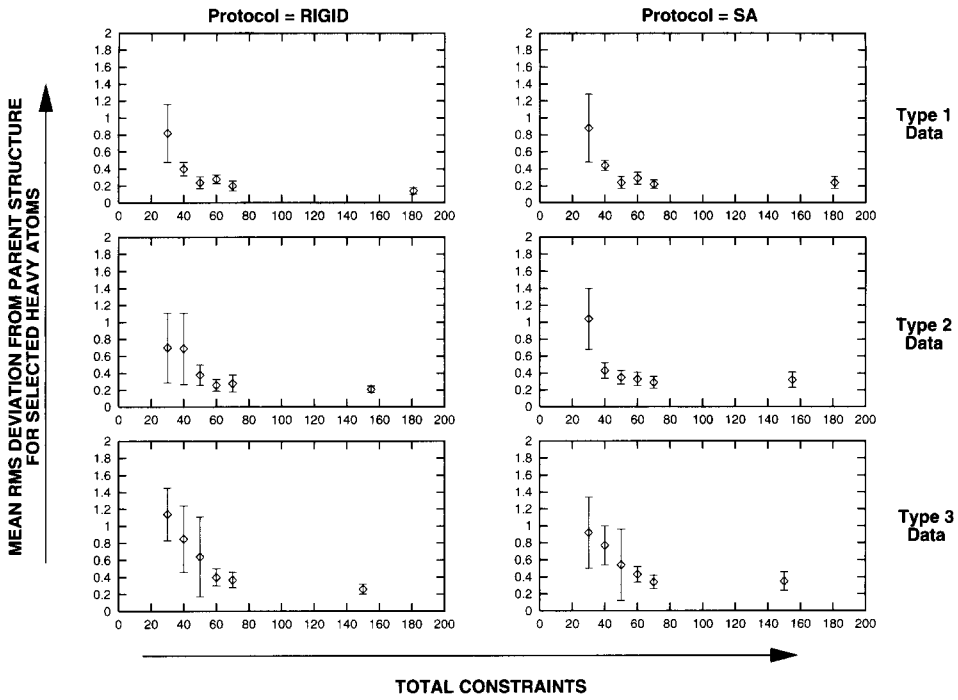


Fig. 4. Variations of the mean rms deviations for sets of 10 calculated docked ligand structures with the number of distance constraints used, for each data type, given an equivalent distribution of simulated NOEs. The quantities were calculated by superimposing each structure with the parent crystal structure and calculating the rms deviation for folate heavy atoms, excluding those of the flexible glutamate moiety. The error bars correspond to the calculated standard deviation. The left column is a plot of the results obtained using the RIGID protocol, whilst the right one is for the SA method. The three rows are for the three different data types (see Fig. 1).

be mistakenly assigned on the basis of an underconstrained docking calculation on a system having a single conformation only. Nevertheless, even when the system is only sparsely constrained, it is still possible to draw some tentative conclusions about the bound ligand orientation.

Effects of NOE constraint averaging methods

A more effective method than the 'geometric centre' (or 'pseudoatom') approach for averaging constraints between groups of protons has been reported (Brünger et al., 1986; Levy et al., 1989; Constantine et al., 1992), based on 'R^{-6}' averaging. The R^{-6}-averaged distance between two groups, 1 and 2, is calculated as:

$$R = (\langle R_{ij}^{-6} \rangle)^{-1/6}$$

where $\langle R_{ij}^{-6} \rangle$ is the mean of the inverse sixth power of the distances between all atoms i in group 1 and j in group 2. If this method is used for averaging of a constraint instead of the pseudoatom approach, it is no longer necessary to add corrections to the upper distance limit, and the restraint is therefore tighter. In this work, modified forms of the original data sets were created that incorporated R^{-6} averaging for NOEs involving protons in methyl groups, NH_2 groups and pairs of symmetrically located aromatic ring protons on the protein and the ligand (except for those cases where the constraint was between averaged nuclei of these types and nonstereospecifically assigned groups of protons). In the parent, type 1, 2 and 3 data sets, 76, 70 and 77 restraints, respectively, utilised R^{-6} averaging. Similar modified forms of the subsets of constraints were also created. New sets of structures were calculated using the RIGID protocol. As above, the overall mean rms deviations for all ligand heavy atoms and for the subset heavy atoms were calculated. The results are presented in Table 2. For large data sets, no improvements in accuracy were found using R^{-6} averaging (in fact, in many cases the small differences seem to favour centre averaging). However, with smaller, more realistic data sets, there were significant improvements on accuracy using R^{-6} averaging, and for very small sets of constraints (~25–35 for the nonglutamate part of the ligand) the improvements were dramatic. A comparison of plots of the mean rms deviations for the subset of ligand atoms described above (i.e., without the glutamate) against the total number of constraints calculated with and without R^{-6} averaging (Fig. 5) clearly demonstrates that the quantity of NOEs required to obtain a single conformation, which is a good representation of the true structure, is reduced by utilising R^{-6} averaging. About 25–35 constraints were sufficient to give a good calculated structure (rmsd < 0.5 Å) for the nonglutamate part of the ligand. This is further clarified by comparing structures calculated using the RIGID protocol with 40 constraints (28 constraints from the nonglutamate part of the ligand),

TABLE 2
MEANS AND STANDARD DEVIATIONS FOR ENSEMBLES OF DOCKED LIGAND STRUCTURES CALCULATED WITH THE RIGID PROTOCOL AND WITH R^{-6} AVERAGING USING THE SAME CONSTRAINT SETS AS IN TABLE 1

Data type	Number of constraints		Mean rms deviation, subset heavy atoms[a] (Å)
	complete ligand	excluding glutamate moiety	
1	181 (All)	123 (All)	0.19±0.01
1	70	50	0.22±0.03
1	60	38	0.30±0.01
1	50	34	0.34±0.03
1	40	28	0.44±0.03
1	30	22	0.42±0.27
2	155 (All)	107 (All)	0.18±0.01
2	70	50	0.34±0.01
2	60	38	0.29±0.01
2	50	34	0.39±0.03
2	40	28	0.40±0.13
2	30	22	0.48±0.21
3	150 (All)	102 (All)	0.30±0.01
3	70	50	0.32±0.01
3	60	38	0.49±0.03
3	50	34	0.44±0.03
3	40	28	0.54±0.02
3	30	22	0.86±0.37

[a] Atoms of the flexible glutamate moiety were not included in the calculation of the rms deviation.

with and without R^{-6} averaging (Fig. 6). In the original families, which use centre averaging for data types 2 and 3, a unique position is not defined for the benzoyl ring. The definition of the pteridine ring is also poor. In comparison, these two rings have unique solutions that are close to the correct structure when R^{-6} averaging is included. Furthermore, for all data types the families of structures are more precisely defined. Clearly, this method improves docking calculations of this type in cases where realistic quantities of constraints are available.

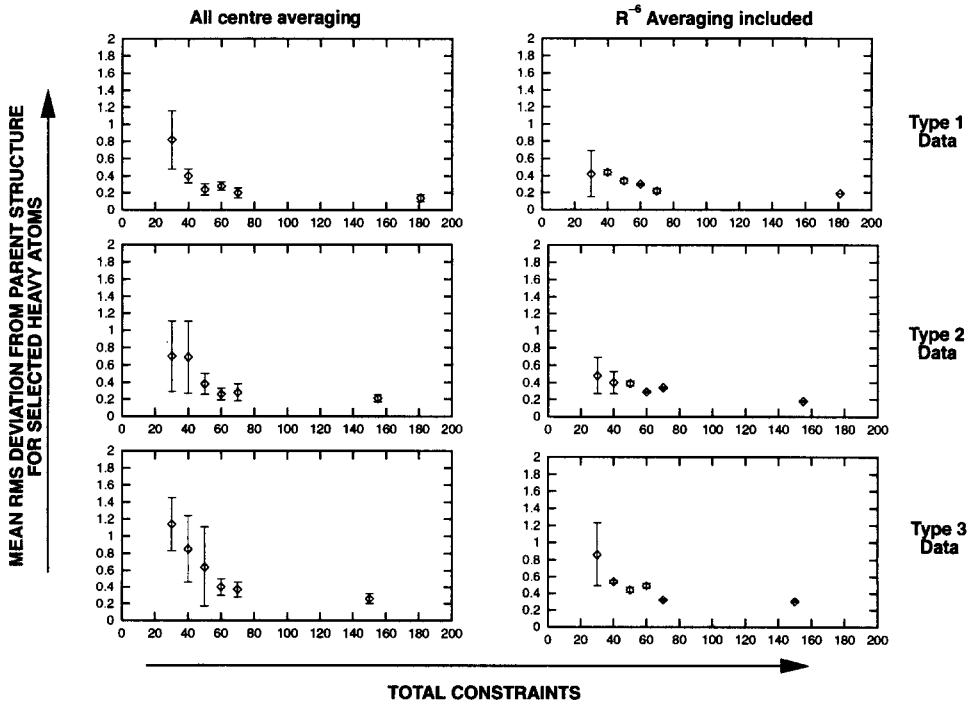


Fig. 5. Variations of the mean rms deviations obtained as in Fig. 4, for structures calculated using the RIGID method. The left column of graphs summarizes results obtained using all centre averaging of the distance constraints, whilst the right one corresponds to the incorporation of R^{-6} averaging.

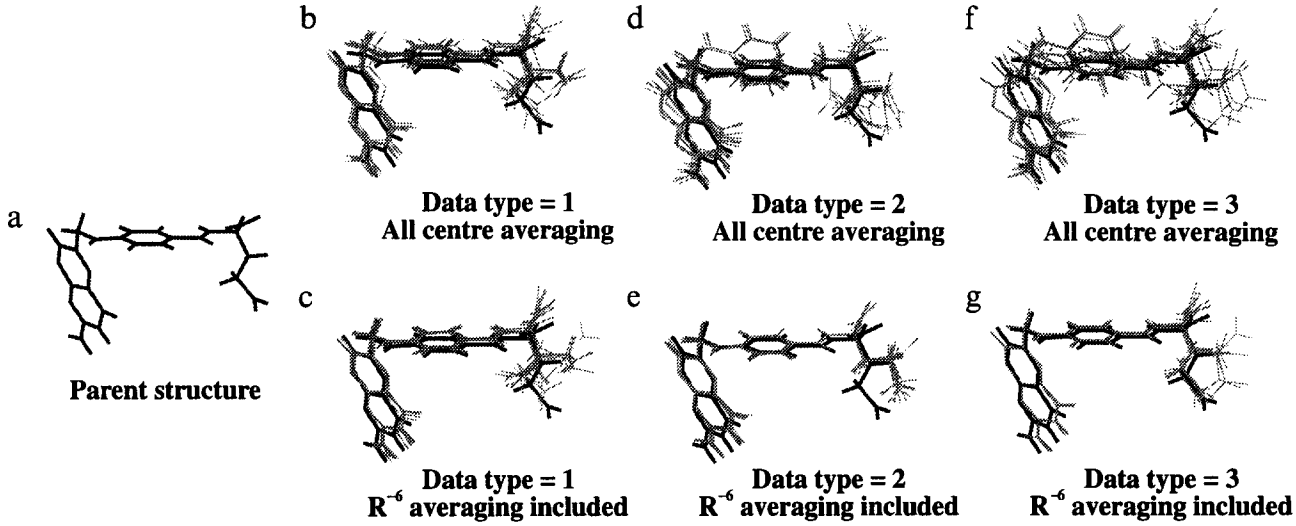


Fig. 6. Sets of structures calculated from RIGID with the 40-constraint series of data sets (28 of the 40 constraints were for the nonglutamate part of the ligand), shown superimposed on the parent structure. In c, e and g, R^{-6} averaging has been used (see the text for details). The original families of structures, with all centre averaging of the constraints, are shown for comparison in b, d and f.

Effects of differences in protein side-chain conformations in different complexes

The overall secondary and tertiary structures of an enzyme in different but related complexes are often very similar. However, the conformations of some side chains, particularly those around the binding site, will usually vary depending on the different bound ligands. In the calculations discussed so far, all the protein atoms were held fixed during the dynamics. In a real docking study, ligand–protein NOE constraints may involve side chains of protein residues that have different conformations compared to those in the initial protein structure used for docking. This will clearly influence the calculated docked ligand conformation, resulting in a degradation in accuracy. In the present study, the effects of such structural changes were assessed by using a modified protein structure in docking calculations. The conformations of the protein side chains were effectively 'scrambled' by performing 5 ps of dynamics at a simulated target temperature of 4000 K on the enzyme alone (the ligand coordinates were deleted), keeping its backbone atoms fixed. Nonbonded interactions were represented using the simplified potential described for the simulated annealing calculations. With the backbone atoms still fixed, the protein was subsequently energy minimized to regularize its geometry. This resulted in a new protein structure, in which 75% of the amino acid residues had a χ_1 torsion angle changed by $>10°$, 60% changed by $>30°$ and 30% changed by $> 60°$ (including several now in different rotameric states). This protein structure was then used as the starting structure for a new series of docking calculations. The subsets with 50 constraints (34 constraints from the nonglutamate part of the ligand) utilised in the earlier studies were used, with R^{-6} averaging, since reasonable structures had been obtained previously using these data sets and they represent a realistic quantity of restraints. In the first approach, no real modification to the existing protocol was used other than increasing the

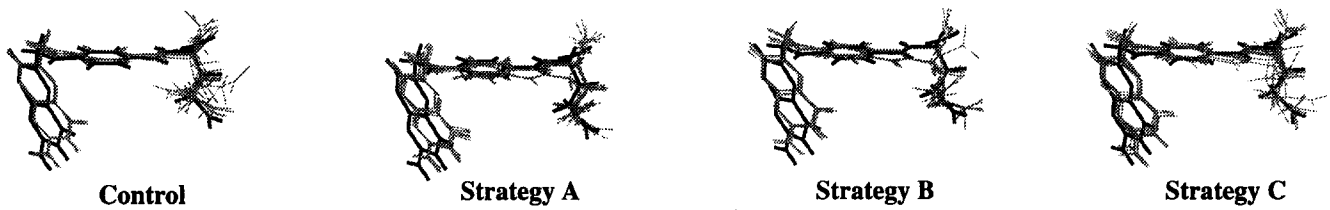


Fig. 7. Sets of structures calculated using the type 3 50-constraint data set (34 of the 50 constraints were for the nonglutamate part of the ligand) and the RIGID protocol with R^{-6} averaging. Each family of calculated structures is superimposed on the parent protein structure as described in previous figure captions. The set of conformations marked 'Control' has been calculated using the unmodified protein structure and the usual RIGID protocol (it corresponds to the 50-constraint set of structures used for the appropriate plot in Fig. 5). The other three sets of structures were calculated using a protein starting structure that had been modified by scrambling the side-chain conformations (see the text for details), and modifying the latter two dynamics stages of RIGID to account for this in one of three ways: In 'strategy A', the original RIGID protocol was used, but with 500 cycles of Powell minimization rather than 200. In 'strategy B', prior to the minimization stage in the RIGID protocol an additional 500 cycles of minimization were performed, with the atoms of the ligand and the protein side chains allowed to move freely. In 'strategy C', the side chains of protein residues with intermolecular NOE constraints were also free to move after the rigid-body dynamics stage of the RIGID protocol.

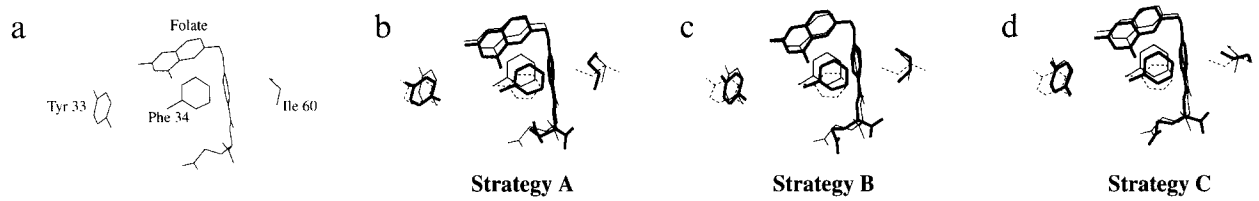


Fig. 8. Diagrams showing the restrained minimized mean structure of the calculated conformations depicted in Fig. 7. In addition to the ligand, the side chains from residues Tyr^{33}, Phe^{34} and Ile^{60} are also shown. (a) Parent structure. (b)–(d) Calculated minimized mean structures are shown in bold type. In each diagram, the modified version of RIGID used (strategy A, B or C) is indicated. The parent structure is shown in thin lines. Side chains from the starting protein structure used are shown in broken lines.

number of steps in the final Powell minimization from 200 to 500. This is referred to as strategy A. Figure 7 shows a set of 10 converged ligand conformations obtained from these calculations superimposed on the parent structure. The original set of 10 structures calculated using these constraints, with the normal X-ray-determined conformation of the protein as the starting structure, are also shown for comparison, described here as the 'control' set. Two modified protocols, termed strategies B and C, were then tried. In the first modification, an initial extra stage was inserted prior to the final minimization, where 500 cycles of Powell minimization were calculated in which ligand and NOE-constrained protein side-chain atoms were allowed to move freely. In strategy C, the side-chain atoms of NOE-constrained protein residues were not held fixed during the 'conventional' molecular dynamics stages. Sets of 10 structures, calculated using both modifications, are shown in Fig. 7. The means and standard deviations of the rms deviation of each family of ligand structures from the parent conformation were also calculated using all heavy atoms in folate or the subset derived by excluding the more flexible glutamate moiety, as described above. These results are presented in Table 3. From inspection of these data and the sets of structures in Fig. 7 it is clear that, although strategy C gives somewhat less precise results than A and B, its calculated structures are more accurate. This is particularly evident in the case of the type 2 and 3 data, where the calculated accuracy is not much worse than that obtained in the 'control'. The reason for this becomes more clear when the constrained protein residues are also accounted for at a qualitative level. This was most easily visualized by calculating restrained minimized mean structures for each family of conformations. The mean structures obtained using each of the three strategies with the type 3 data set are shown superimposed on the original crystal structure and the modified enzyme coordinates in Fig. 8. Side chains from three sample residues in the protein, Tyr^{33}, Phe^{34} and Ile^{60}, are also shown. These were chosen as appropriate examples due to the fact that methyl groups and aromatic ring protons have easily identifiable NMR signals. The results suggest that strategy C, where dynamics are performed on the constrained protein side chains, allowing quite efficient sampling of the conformational space available to these atoms, is the most successful at predicting the correct conformation and configuration of these residues and subsequently obtaining more accurate structural parameters for the bound ligand. For this data set, it performs particularly better in positioning the aromatic rings. Thus, incorporation of essentially all constrained atoms into the dynamics stage of the docking gave the best results.

TABLE 3
MEANS AND STANDARD DEVIATIONS FOR ENSEMBLES OF DOCKED LIGAND STRUCTURES CALCULATED WITH THE 50-CONSTRAINT, R^{-6}-AVERAGED DATA SETS

Data type	Calculation strategy[a]	Mean rms deviation, subset heavy atoms[b] (Å)
1	Control	0.34±0.03
1	A	0.52±0.04
1	B	0.55±0.04
1	C	0.51±0.19
2	Control	0.39±0.03
2	A	0.67±0.05
2	B	0.72±0.03
2	C	0.58±0.08
3	Control	0.44±0.03
3	A	0.72±0.03
3	B	0.84±0.18
3	C	0.66±0.06

The data sets contained 34 of the 50 constraints for the nonglutamate part of the ligand. Where indicated, the modified protein structure was used (see text), and different modifications to the rigid protocol were tested, as described under 'calculation strategy'.

[a] Control refers to the original structures obtained using the unmodified protein structure. Calculation strategies A, B and C refer to three modified versions of RIGID, using the protein structure with 'scrambled' side chains as the starting structure (see the text for details).

[b] Atoms of the flexible glutamate moiety were not included in the calculation of the rms deviation.

Conclusions

From this investigation of the effectiveness and methodology of NOE docking, several conclusions can be drawn. (i) The results indicate that the approach adopted here provides a valid method for obtaining structural information for a complex of a protein when a high-resol-

ution structure of the protein in a related complex is already available. (ii) The simulated annealing protocol in XPLOR 3.1 (Brünger, 1992; 'sa.inp'), used with fixed coordinates for the unconstrained parts of the protein, provides a good method for the NOE-constrained docking. However, a rigid-body dynamics method has also been shown to be a fast and easily implementable way of locating the ligand in its correct binding pocket, with additional refinement to obtain the correct conformation, as proposed by Yue (1990). (iii) In cases where small, realistic, NOE-constrained data sets were used, R^{-6} averaging rather than centre averaging gives large improvements in the accuracy of the calculated structures. (iv) In NOE-constrained docking studies, improved accuracy of the results is achieved by allowing the NOE-constrained regions of the protein to move during dynamics. (v) The dependence of the precision and accuracy of the structures on the number of constraints used has been found to be more severe for the NOE-based docking than for complete protein structure determination (Clore et al., 1986,1993; Liu et al., 1992; Zhao and Jardetzky, 1994). This is because the conformation and the orientational freedom of the ligand are constrained only by the NOEs and not by covalent attachment to the protein, which gives some increase in freedom compared with an amino acid side chain in the enzyme. In NMR structure determination of proteins, the number of experimental constraints per residue required to provide accurate structures was found to be approximately 15 (Clore et al., 1993). For most amino acids, this corresponds to less than eight constraints per torsion angle. The number of constraints required to give a good structure for the part of folate encompassing the two rings (which contains three torsion angles) is in the range 25–35 if R^{-6} averaging is used. This provides a useful guide for estimating the constraint requirements for work in our laboratory on the NOE docking of antifolate drugs (Martorell et al., 1994; Morgan et al., 1995).

Some of the conclusions of this work can be extended to complete NMR structure determination of macromolecule–ligand complexes, in terms of the definition of the ligand in calculated structures. As above, it will be advantageous to use R^{-6} averaging. Rigid-body docking could also be usefully incorporated into the total structure determination protocol; after calculating a first generation structure for the protein, the ligand may be introduced in this way prior to subsequent iterative refinements.

Acknowledgements

We wish to thank András Aszódi, Berry Birdsall, Angelo Gargaro, Andrew N. Lane, William D. Morgan, Vladimir I. Polshakov and William Taylor for helpful discussions. M.J.G. acknowledges the award of an M.R.C. research fellowship.

References

Bennion, C., Connolly, S., Gensmantel, N.P., Hallam, C., Jackson, C.G., Primrose, W.U., Roberts, G.C.K., Robinson, D.H. and Slaich, P.K. (1992) *J. Med. Chem.*, **35**, 2939–2951.

Billeter, M., Havel, T.F. and Kuntz, I.D. (1987) *Biopolymers*, **26**, 777–793.

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

Brünger, A.T. and Karplus, M. (1988) *Protein Struct. Funct. Genet.*, **4**, 148–156.

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

Byeon, I.L., Kelley, R.F., Mulkerrin, M.G., An, S.S.A. and Llinás, M. (1995) *Biochemistry*, **34**, 2739–2750.

Clore, G.M., Brünger, A.T., Karplus, M. and Gronenborn, A.M. (1986) *J. Mol. Biol.*, **191**, 523–551.

Clore, G.M., Robien, M.A. and Gronenborn, A.M. (1993) *J. Mol. Biol.*, **231**, 82–102.

Constantine, K.L., Madrid, M., Bányai, L., Trexler, M., Patthy, L. and Llinás, M. (1992) *J. Mol. Biol.*, **223**, 281–298.

Fesik, S.W., Neri, P., Meadows, R., Olejniczak, E.T. and Gemmecker, G. (1992) *J. Am. Chem. Soc.*, **114**, 3165–3166.

Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. (1983) *Science*, **220**, 671–680.

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

Lian, L.Y., Barsukov, I.L., Sutcliffe, M.J., Sze, K.H. and Roberts, G.C.K. (1994) *Methods Enzymol.*, **239**, 657–700.

Liu, Y., Zhao, D., Altman, R. and Jardetzky, O. (1992) *J. Biomol. NMR*, **2**, 373–388.

Martorell, G., Gradwell, M.J., Birdsall, B., Bauer, C.J., Frenkiel, T.A., Cheung, H.T.A., Polshakov, V.I., Kuyper, L. and Feeney, J. (1994) *Biochemistry*, **33**, 12416–12426.

Meadows, R.P. and Hajduk, P.J. (1995) *J. Biomol. NMR*, **6**, 41–47.

Metropolis, N., Rosenbluth, M., Rosenbluth, A., Teller, A. and Teller, E. (1953) *J. Chem. Phys.*, **21**, 1087–1092.

Morgan, W.D., Birdsall, B., Polshakov, V.I., Šali, D., Kompis, I. and Feeney, J. (1995) *Biochemistry*, **34**, 11690–11702.

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

Oefner, C., D'Arcy, A. and Winkler, F.K. (1988) *Eur. J. Biochem.*, **174**, 377–385.

Powell, M.J.D. (1977) *Math. Program.*, **12**, 241–254.

Rykaert, J.P., Ciccotti, G. and Berendsen, H.J.C. (1977) *J. Comput. Phys.*, **23**, 327–341.

Weber, D.J., Gittis, A.G., Mullen, G.P., Abeygunawardana, C., Lattman, E.E. and Mildvan, A.S. (1992) *Protein Struct. Funct. Genet.*, **13**, 275–287.

Weber, D.J., Serpersu, E.H., Gittis, A.G., Lattman, E.E. and Mildvan, A.S. (1993) *Protein Struct. Funct. Genet.*, **17**, 20–35.

Weber, D.J., Libson, A.M., Gittis, A.G., Lebowitz, M.S. and Mildvan, A.S. (1994) *Biochemistry*, **33**, 8017–8028.

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

Yue, S. (1990) *Protein Eng.*, **4**, 177–184.

Zhao, D. and Jardetzky, O. (1994) *J. Mol. Biol.*, **239**, 601–607.