Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

NMR trial models: experiences with the colicin immunity protein Im7 and the p85α C-terminal SH2–peptide complex

Richard A. Pauptit,[a]* Caitríona A. Dennis,[b] Dean J. Derbyshire,[c] Alexander L. Breeze,[a] Simon A. Weston,[a] Siân Rowsell[a] and Garib N. Murshudov[d]

[a]AstraZeneca, Mereside, Alderley Park, Macclesfield, Cheshire SK10 4TG, England, [b]Astbury Centre for Structural Molecular Biology, Leeds University, Leeds LS2 9JT, England, [c]Department of Haematology, The Wellcome Trust Centre for Molecular Mechanisms in Disease, University of Cambridge, Hills Road, Cambridge CB2 2XY, England, and [d]Department of Chemistry, University of York, York YO10 5DD, England

Correspondence e-mail: richard.pauptit@astrazeneca.com

Received 12 March 2001
Accepted 23 May 2001

PDB References: human p85α C-terminal SH2–peptide complex, 1h9o; Im7, 1ayi.

Two cases of successful molecular replacement using NMR trial models are presented. One is the crystal structure of the *Escherichia coli* colicin immunity protein Im7; the other is a heretofore unreported crystal structure of a specific PDGF receptor-derived peptide complex of the carboxy-terminal SH2 domain from the p85α subunit of human phosphatidylinositol 3-OH kinase. In both cases, molecular replacement was non-trivial. Success was achieved using trial models that consisted of an ensemble of NMR structures from which the more flexible portions had been excized. Use of maximum-likelihood refinement proved critical to be able to refine the poor starting models. The challenges typical of the use of NMR trial models in molecular replacement are discussed.

1. Introduction

In macromolecular crystallography, a crystal structure may be solved by molecular replacement, *i.e.* by orienting and positioning a related trial structure in the unknown crystal unit cell, if a structure for a sufficiently closely related trial model is available. CCP4 Study Weekends on molecular replacement have taken place in 1985, 1992 and now in 2001. Comparison of the state of the field at each of these provides an opportunity to monitor any progress. In molecular replacement, successful use of a trial model that was determined by NMR spectroscopy was unheard of in 1985, under consideration in 1992 and, over the last decade, put to practice in at least 23 applications (Chen *et al.*, 2000; Chen, 2001). What has made this possible? A trivial answer is that the bulk of NMR structures available in the Protein Data Bank (PDB; Berman *et al.*, 2000) have been deposited within the last decade. There are now over 2000 NMR entries in the PDB, representing a significant potential phasing resource for crystallographers.

Unfortunately, NMR models are reputedly a particularly difficult starting point for crystallographic phasing. Long-range order may not be well defined in NMR models. The NMR model represents the solution state of the macromolecule, which may be associated with a greater conformational flexibility such that atomic positions are not as precisely defined as in a target solid-state crystal structure. This suggests that it would be more difficult to detect overlap between the observed (from crystallographic diffraction data) and calculated (from NMR model coordinates) Patterson functions. Furthermore, NMR models are typically of small proteins, which complicates molecular replacement in a number of ways. There are fewer self-vectors relative to the number of cross-vectors, thus the rotation function is noisier. There is less

of a protein core, hence a relatively larger portion of the molecule is closer to the surface, with implications for flexibility: one does not have to move far from the centre of the molecule before reaching loop structures, which may be ill-defined or simply different between trial and target structures. Also, small domains or proteins are often nearly spherical and less amenable to packing discrimination for incorrect positions than are multidomain proteins with defined anisotropic shape. It is possible to rotate a correctly positioned small spherical molecule over a large number of angular permutations without clashing with neighbouring molecules. This renders one of the most useful discriminators for a correct molecular-replacement solution, that of acceptable crystal packing, much less effective. Still, success has been obtained in a significant number of cases in the last decade. What has allowed this? Is there an improved molecular-replacement methodology which now accommodates the use of NMR trial models? In an attempt to address this and arrive at possible generalizations, we examine two successful structure determinations from our laboratory where NMR models were used to obtain phases *via* molecular replacement.

2. Case 1: the colicin immunity protein Im7 from *E. coli*

2.1. Biological background

Colicins are toxins secreted by *E. coli* under conditions of stress. They are classified into groups according to the cell-surface receptor they bind on the target cells, *e.g.* E colicins bind the BtuB receptor (James *et al.*, 1996). E colicins have three domains: a receptor-binding domain, a translocation domain and a cytotoxic domain. To protect the producing cell, a 9.5 kDa (around 90 amino-acid residues) immunity protein which inhibits the cytotoxic domain is co-expressed with the colicin. The colicin and immunity protein form an extremely tight complex ($K_d \simeq 10^{-17}\ M$) which is secreted and binds to the target-cell membrane receptor. Upon internalization in the target cell, the complex dissociates, thereby activating the toxic domain which kills the cell. The structure determined here is that of Im7, the immunity protein specific to colicin E7, which is a DNase.

2.2. Experimental

Im7 was purified as described by Wallis and coworkers (Wallis, Leung *et al.*, 1995; Wallis, Moore *et al.*, 1995). The structure and its determination are described in detail by Dennis *et al.* (1998); relevant details are summarized here. Crystals were either *I*222 or $I2_12_12_1$ (the translation function worked only in *I*222; see below), with unit-cell parameters $a = 45.1$, $b = 50.6$, $c = 75.2$ Å, $\alpha = \beta = \gamma = 90^\circ$. Data were 93% complete to 2.0 Å, with an $R_{\rm sym}$ of 4.9%.

The molecular replacement was difficult and time-consuming. When this project commenced, the only related structure that was available was an NMR structure of Im9 (Osborne *et al.*, 1996). Im9 is the cognate immunity protein for colicin E9. Im9 shares only 60% sequence homology with Im7, generating considerable concern for the success of molecular replacement: all previously reported successes with the use of NMR trial models were for cases where there was 100% homology between trial and target structure, *i.e.* the X-ray structure was determined using an NMR model for the identical protein. The problem was taken to a week-long CCP4 workshop on molecular replacement in York in 1996 and benefited from in-depth discussion. The factors that led to success are as follows. A trial model was generated from the superposed ensemble of the ten lowest energy NMR structures (all trials carried out using a single NMR structure yielded no correct answers). Regions of extreme main-chain flexibility at the N- and C-termini and some loops (residues 1–5, 25–31, 56–62 and 82–86) were removed from the model since these were considered regions of undefined structure which would add noise rather than signal to the molecular replacement. Non-conserved side chains were replaced by alanine, as were conserved side chains that showed differing conformations within the ensemble. The atoms in the model were all assigned a temperature factor of 20 Å^2, a convenient estimate calculated from the average r.m.s.d. (0.53 Å) for the backbone atom positions in the ensemble. The data were limited to the resolution range 15–4.5 Å (completeness 95%) to approximate a high-resolution limit at which the trial and target structures might be considered to be the same. The integration radius was restricted to 16 Å. The final model contained only 60% of all the atoms, but included the secondary-structure elements (four helices).

Using the *CCP*4 program *AMoRe* (Navaza, 1994; Collaborative Computational Project, Number 4, 1994), no discrimination could be observed in the rotation function. The 100 highest peaks were considered for the translation function, which was carried out in both *I*222 and $I2_12_12_1$ since the space-group ambiguity needed to be resolved. The highest translation function peak in space group *I*222 proved to be correct, though it did not stand out significantly from the background. The correlation coefficient (0.54) was 0.02 higher than the highest noise peak; the *R* factor (50%), though the lowest, was not significantly lower than the noise level. The rotation-function peak for which the correct translation was found ranked 57th in the list of peaks obtained in the rotation search. Rigid-body refinement (also at 4.5 Å resolution) of the translation-function solution increased the correlation coefficient to 0.56 (0.04 higher than the next peak) and lowered the *R* factor to 48% (2% lower than the next peak). The crystal packing for this solution showed a tight molecular packing with no steric violations; $2F_o - F_c$ difference electron density could be observed for some of the side chains that had been trimmed to alanine residues. The latter was taken as absolute confirmation of correct molecular replacement.

A starting model for refinement was created by selection of the NMR model which was closest to the average of the ensemble. This was considered preferable to using the average structure, which might be thought artificial. The average structure may satisfy experimental data as it is usually derived by taking a coordinate average and energy minimizing with the restraint potentials switched on. However, this average is not truly representative of an experimental structure, since it

has not been found by restrained simulated annealing or distance geometry starting from a random structure. The starting model for refinement was incomplete owing to the excized portions of the molecule and also not as representative of the true structure as was the ensemble, hence the initial derived phases were poor. Simulated-annealing refinement (no maximum-likelihood target) with *X-PLOR* (Brünger, 1992) for several cycles interspersed with manual rebuilding using *O* (Jones *et al.*, 1991) reduced R and $R_{\rm free}$ (using 10% of data) from 50 and 50%, respectively, to 39 and 48% using 8–3 Å data, and then to 37 and 46% using 8–2 Å data. Some side chains and some of the missing loop residues were built into weak difference density. Conventional positional refinement could reduce R to 31%, but $R_{\rm free}$ remained high (43%) and refinement progressed no further. The high $R_{\rm free}$ was interpreted as overfitting. To delete any incorrectly built regions, the loop residues that had been built in were removed and the maximum-likelihood refinement program *REFMAC* (Murshudov *et al.*, 1997) from the *CCP*4 program suite (Collaborative Computational Project, Number 4, 1994) was then applied to the incomplete model, resulting in superior difference density. Ten cycles reduced R to 33% using 10–2 Å data. The entire loop structure could be built and refinement continued to R = 21.2%, $R_{\rm free}$ = 26.5%, at which stage water molecules could be added. Refinement converged with R = 17.8% and $R_{\rm free}$ = 24.7% using data in the resolution range 10–2 Å.

2.3. Description of structure

The structure is a distorted four-helical bundle, as described in detail in Dennis *et al.* (1998), where it is compared with the structure of the related Im9 (Osborne *et al.*, 1996) to account for differences in binding specificities to their cognate colicin DNase domains. It is worth mentioning that subsequent studies on the crystal structures of the complex between Im7 with the DNase domain of colicin E7 (Ko *et al.*, 1999) as well as the complex of Im9 with the DNase domain of colicin E9 (Kleanthous *et al.*, 1999) have shown that the immunity protein remarkably does not bind at the highly conserved DNase active site. This allows exploitation of interactions with the more variable portions of the DNase to achieve specifity. Inhibition is achieved by steric hindrance and electrostatic repulsion afforded by the bound immunity protein which does not allow the approach of the substrate, bacterial DNA, a large polymer which extends well beyond the active site (Kleanthous *et al.*, 1999).

3. Case 2: specific peptide complex of the p85α C-terminal SH2 domain from human phosphatidylinositol 3-OH kinase

3.1. Biological background

SH2 (src-homology 2) domains are small protein modules of around 100 amino acids that bind phosphotyrosine moieties in a selective amino-acid sequence context (Pawson & Schlessinger, 1993). SH2 domains play a key recognition role in a large number of signal-transduction pathways. Type 1A phosphotidylinositol 3-OH kinase (PI3 kinase-α) is a lipid kinase whose phosphorylated lipid products mediate cytoplasmic signalling pathways in response to extracellular stimuli such as growth-factor-induced mitogenic responses (Kapeller & Cantley, 1994). Intervention is perceived to have possible therapeutic applications in cancer. PI3 kinase-α is a heterodimer, consisting of a 110 kDa (p110) catalytic subunit and an 85 kDa regulatory subunit (p85α). The p85α subunit contains an src-homology 3 (SH3) domain, a breakpoint cluster (BCR) domain and then two SH2 domains separated by a 200-residue linker which binds and activates the p110 kinase subunit. The two SH2 domains recognize similar phosphotyrosine-containing consensus motifs on the cytoplasmic tails of activated growth factor receptor tyrosine kinases. Specificity is incurred through a strong preference, especially in the C-terminal SH2 domain studied here, of a methionine residue in the $i + 3$ position relative to the phosphotyrosine (Songyang *et al.*, 1993). The complex studied here is that with the pentapeptide pTyr-Val-Pro-Met-Leu, which is the specificity sequence around the Tyr751 binding site on the platelet-derived growth factor (PDGF) receptor. This receptor becomes autophosphorylated on extracellular binding of ligand, thereby allowing binding by the p85α SH2 domains of PI3-kinase, resulting in activation of the kinase and production of phosphorylated lipid signalling.

The crystal structure of the SH2 domain–ligand complex was undertaken in order to assist design of a phosphotyrosine mimic that would interfere with the productive association of PI3-kinase with activated PDGF receptor. A substantial number of SH2 structures, determined by both NMR and X-ray crystallography, are available in the PDB and represent potential trial models. For the p85α C-terminal SH2 domain, an NMR model for the human protein complexed to the same pentapeptide (albeit acetylated) has been elucidated at AstraZeneca (Breeze *et al.*, 1996; PDB accession code 1pic), and, more recently, an X-ray structure for bovine p85α C-SH2 in the absence of peptide has been published (Hoedemaeker *et al.*, 1999; PDB code 1qad). Unfortunately, the latter structure was not completed at the time this work was undertaken. There is also an NMR structure available for bovine p85α N-terminal SH2 model in the absence of bound peptide (Booker *et al.*, 1992; PDB code 2pnb), of which a crystal structure was later determined for the human protein, both in the presence and absence of peptide ligands (Nolte *et al.*, 1996).

3.2. Experimental

Protein and peptide were prepared as described in Breeze *et al.* (1996). Only a few good crystals were obtained, precluding the use of MIR methods for structure solution. Crystallization by hanging-drop vapour diffusion was hampered by limited reproducibility, compounded by the fact that the few best crystals obtained had grown in experiments with inadvertently cracked cover slips. The crystals that did grow belonged to space group $C2$, with unit-cell parameters a = 59.0, b = 32.8,

c = 54.9 Å, β = 96.2°. Data were collected on a 30 cm MAR image plate mounted on an Enraf–Nonius FR571 rotating-anode X-ray generator operating at 40 kV and 90 mA. Each image was 1° of crystal rotation; 200 images were collected at room temperature with a crystal-to-detector distance of 120 mm. Data were processed using *XDS* software (Kabsch, 1993): there were 9283 unique reflections from 36 576 observations, corresponding to a 93% complete data set to 1.79 Å. The data merged with an overall $R_{\rm sym}$ of 5.0%.

Despite enormous effort, all attempts to solve the structure *via* molecular replacement using as trial model any of the X-ray crystal structures of SH2 domains available in the PDB failed. Sequence identity between different SH2 domains can be as high as 50% (*e.g.* between src and lck SH2). The p85α C-terminal SH2, however, shows little more than 20% identity with the sequences of available crystal structures. The exception is the human p85α N-terminal SH2 (Nolte *et al.*, 1996), with which 35% identity is shared; however, this structure was unavailable at the start of this project and at the present time coordinates are not deposited in the PDB.

At the same time that the p85α C-terminal SH2–peptide complex underwent crystallization trials, the solution study by NMR was initiated (Breeze *et al.*, 1996). An intermediate NMR model (calculated before all NOE, torsion angle and hydrogen-bond restraints were included) became available and could be used for molecular replacement with the program *AMoRe* (Navaza, 1994; Collaborative Computational Project, Number 4, 1994). In this case, there is 100% homology between trial and target proteins. Using an ensemble of the ten lowest energy structures, in which poorly defined side chains and portions of the main chain had been excized and a temperature factor of 30 Å^2 was assigned to all remaining atoms, the highest peaks in a rotation function and translation function calculated to 4.5 Å (data are 98% complete at this resolution) corresponded to an acceptable packing arrangement. The integration radius for the rotation function was 16 Å. Two rotation functions solutions of nearly equal height were always obtained, corresponding to the two possible orientations of the molecule in the $C2$ cell: choosing one or the other corresponds to a choice of origin. Convincingly, the rotation-function value of the two highest peaks increased as additional NOE restraints were used to generate more accurate NMR models. Thus, in four stages of increased accuracy in the NMR model, the molecular replacement solution was progressively more visible. At each stage, the NMR models were trimmed in a consistent manner (use of a script to excize flexible portions ensured this). This improvement in the rotation-function value as the accuracy of the trial model increased was taken as evidence of a genuine molecular replacement solution. With the final NMR model, the two highest rotation-function correlation coefficients were 0.32 and 0.29, with the next highest value in the rotation-function map being 0.21 (corresponding to the noise level). The translation function gave an R factor of 51% and correlation coefficient of 0.25. This improved to 47% and 0.33, respectively, on rigid-body refinement (at 4.5 Å). Unit-cell packing was acceptable, but initial electron-density maps were barely interpretable and gave no clear indications of new information for side-chain orientation or missing loops. However, there was clear density for the phosphotyrosine which had been omitted from the map calculation in order that it may be used as a marker. Despite the poor maps, this was taken as further evidence that the molecular replacement was correct.

A single NMR model, that closest to the average structure, was selected from the ensemble for the purpose of refinement. Attempts to refine the model using the slow-cooling simulated-annealing protocol (not with maximum-likelihood targets) available with the program *X-PLOR* (Brünger, 1992) proved unproductive. After five cycles of slow cooling, positional refinement and temperature-factor refinement interspersed with attempts at manual rebuilding, the R factor would decrease to 32%, but $R_{\rm free}$ remained above 50% and the electron-density maps showed no signs of improvement. This happened regardless of resolution range, choice of initial model *etc*. We sent our diffraction data and molecular-replacement solution to University College, London, where an attempt was made by Mark Roe and Laurence Pearl (data not shown) at cross-crystal averaging using their data for orthorhombic ($P2_12_12_1$) unliganded bovine crystals that were subsequently solved independently (Hoedemaeker *et al.*, 1999). In agreement with our own difficulties, these researchers had, at the time, managed to obtain a molecular-replacement solution through use of an NMR trial model but were unable to refine it successfully. A cross-rotation function solution relating the two diffraction data sets was readily obtained and corresponded to the transformation between the two sets of coordinates obtained *via* molecular replacement. This is not absolute proof but suggests strongly that both molecular-replacement solutions are genuine. However,

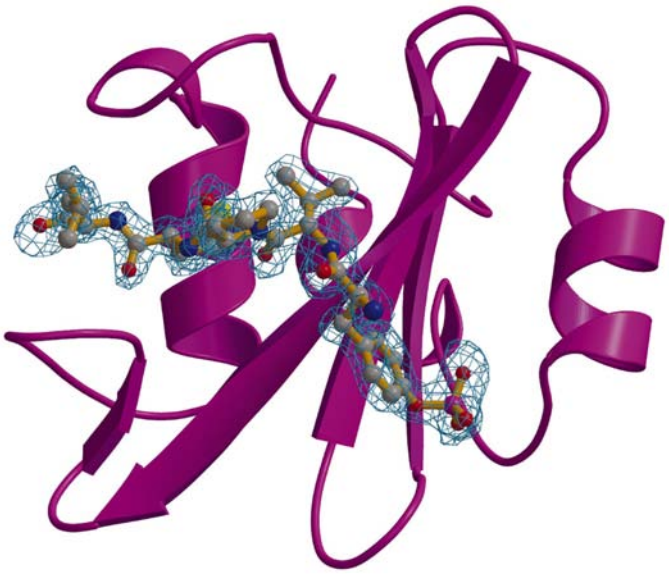

Figure 1
Ribbon drawing of the p85α C-terminal SH2 with the pentapeptide in $2F_o - F_c$ density, generated using *Bobscript* (Esnouf, 1997) and *Raster3D* (Bacon & Anderson, 1998).

subsequent averaging did not result in a more interpretable electron-density map.

The project was abandoned until it was taken to a CCP4 one-week refinement workshop in York, 1998 as a 'difficult' case. There, a number of refinement protocols were tested and success was finally obtained using the *CCP*4 program *REFMAC* (Murshudov *et al.*, 1997; Collaborative Computational Project, Number 4, 1994) in conjunction with *ARP*/*wARP* (Perrakis *et al.*, 1999; Collaborative Computational Project, Number 4, 1994), using a maximum-likelihood target function. Restrained *REFMAC* was used with van der Waals repulsions removed. In *ARP*/*wARP*, a minimum distance of 1.2 Å between old and new atoms was used. The free *R* value decreased to 38%, at which stage the electron-density difference maps were interpretable. This was repeated for several cycles with rebuilding with *QUANTA* (MSI; San Diego, California, USA) until the model was complete, whereupon *REFMAC* was run in the usual fashion including van der Waals restraints. Water molecules were added with the *QUANTA* Xsolvate option. The final *R* factor is 16.3% ($R_{\rm free}$ = 21.9%) and the model has good stereochemistry (r.m.s. bond distances 0.018 Å; r.m.s. bond angles 2.07°).

3.3. Description of structure

It is of interest to compare the structure to other SH2 domains that might have been useful trial models. In particular, it is interesting to compare in some detail with the NMR trial model that was used in order to see if the differences might account for the difficult molecular replacement. The p85α C-terminal SH2 structure determined here is typical of SH2 domains, with a three-stranded β-sheet sandwiched between two α-helices (Fig. 1). With respect to src SH2 (coordinates from our own work, data not shown), there is an eight-residue N-terminal extension which is oriented differently and there are substantial differences in the loop orientations between residues 92 and 99 and between residues 52 and 54 where src has a larger loop. The core of the molecule, however, is quite similar to the point where one might have expected that a trimmed src model would work in molecular replacement despite the low homology. The r.m.s.d. of 65 equivalent C^{α} atoms is 0.9 Å. The structure is extremely similar to that of the unliganded bovine p85α C-terminal SH2 (Hoedemaeker *et al.*, 1999; PDB code 1qad). Here, the r.m.s.d of 99 equivalent C^{α} atoms is 0.5 Å and many side chains have identical conformations. The only noteworthy difference is that the loop 39–44 has become ordered in our structure and interacts with the phosphotyrosine in the pentapeptide ligand, which is of course absent in the bovine structure. When compared with the minimized mean NMR structure of the identical protein (Breeze *et al.*, 1996; PDB code 1pic), *i.e.* the trial model, the structure is again extremely similar for the core, while two loop areas differ: the phosphotyrosine-binding loop (residues 39–44) comes about 4 Å closer to the pentapeptide ligand in the X-ray structure, while the region 50–77 shows significant shifts in the same

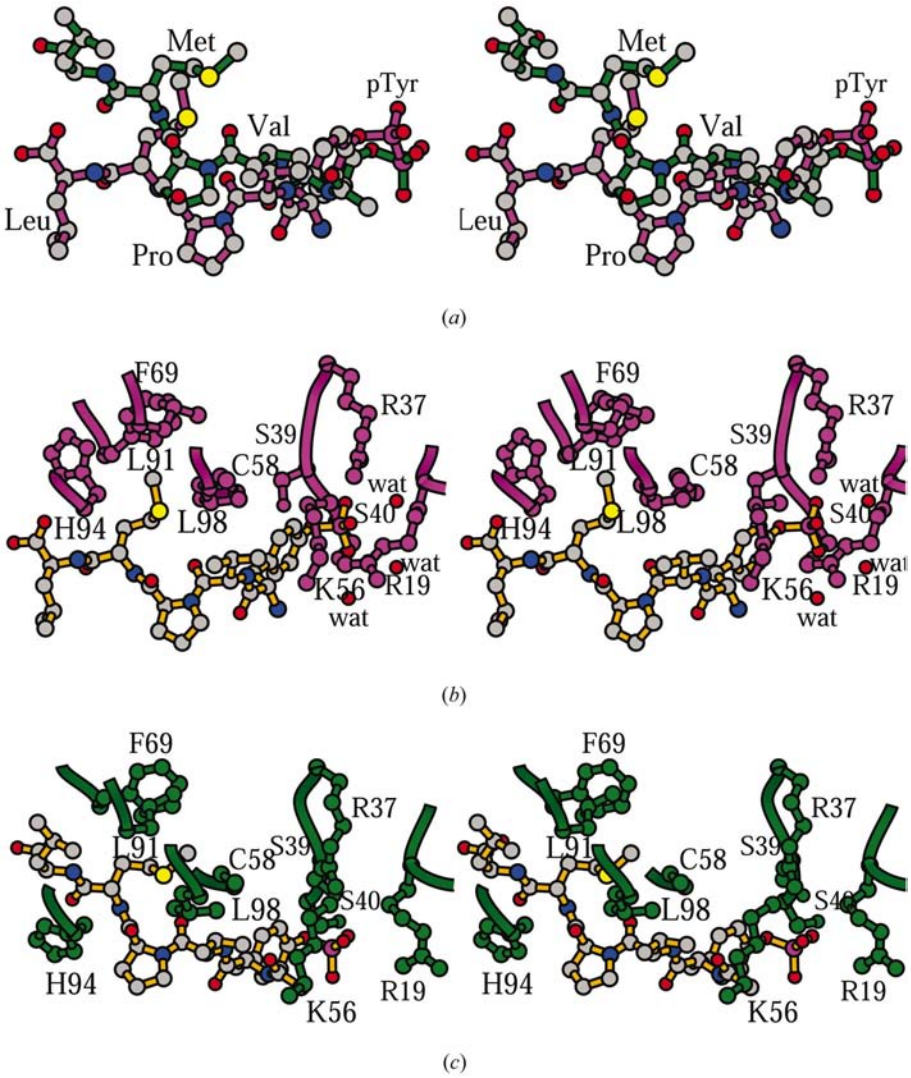

Figure 2
A closer stereoview of the binding of the pentapeptide ligand in the crystal structure (magenta) and in the NMR model (green). (*a*) The relative position of the peptide in both structures when the SH2 domains are superposed, (*b*) the environment in the crystal structure and (*c*) the environment in the NMR structure, showing neighbouring residues mentioned in the text. The figure was generated using *Bobscript* (Esnouf, 1997).

direction. In this region there is also some secondary structure in the crystal not apparent in the NMR structure. Residues 61–63 in the crystal structure form β-sheet interactions with residues 66–68, with a β-turn at residues 64 and 65; in contrast, in the solution NMR studies resonances in this region were severely broadened or absent from the spectra, suggesting mobility on a microsecond to millisecond timescale, and consequently few experimental structural restraints were derived from the data. The N- and C-termini also display differences: the N-termini are oriented in opposite directions and in the crystal the C-terminus is disordered beyond residue 108. This is likely to reflect the arbitrary orientations of these mobile terminal regions, which are largely unrestrained by experimental data in the NMR structure. Further, smaller differences include residues 17 and 18, which adopt different main-chain orientations, as well as the loop 93–98, which is slightly shifted (2 Å) towards the ligand-binding pocket. The r.m.s.d. for 82 equivalent C^{α} positions is 1.0 Å; if 106 C^{α} positions are used, the r.m.s.d. increases to 1.9 Å.

The most remarkable difference between the NMR and the X-ray structure concerns the position of the phosphorylated pentapeptide ligand (Fig. 2). Electron density for the peptide is absolutely clear, suggesting it is highly ordered, yet it appears in the X-ray structure that the ligand is more loosely bound in the binding pocket, *i.e.* the ligand C-terminus is about 5 Å further out of the pocket than in the NMR structure. This movement is in tandem with the change in loop orientation around residue 70 and with a different side-chain orientation for His94. These differences might reflect true structural variations between the crystal and solution environments; it is also possible that the observed NMR conformation is influenced by the necessity to satisfy intermolecular NOEs both to well defined regions of the protein core and to one or two key side chains (for example Phe69) in the otherwise poorly restrained region 50–77 (see above). In the crystal structure, the methionine residue at position 4 in the pentapeptide (the specificity methionine at position $i + 3$ with respect to the phosphotyrosine) is not as deep in the hydrophobic specificity pocket as in the NMR structure or as deep as it could be in the X-ray structure; there is additional space available in the pocket. In the NMR model the methionine is positioned 3 Å deeper in the pocket (Fig. 2). This suggests that the methionine position in the NMR structure may be the more relevant biologically. This hydrophobic pocket is formed by residues Phe69, Cys58, Leu98, Leu91 and His94. Also, the valine residue at position 2 in the pentapeptide could be deeper in the hydrophobic pocket created by Leu98, Cys58 and Lys56. Here, however, there is only 1 Å difference between the X-ray and NMR positions. While there is a 1.3 Å discrepancy between the NMR and X-ray phosphotyrosine C^{α} positions, the tyrosine ring occupies the same space in the phosphate-binding pocket. However, the phosphate group is shifted and the phosphate hydrogen bonding differs: in the crystal, O4 (corresponding to the tyrosine hydroxyl group) is hydrogen bonded to Ser39, O1 is hydrogen bonded to Arg19 which has moved closer with respect to the NMR structure as well as to Arg37, O2 is hydrogen bonded to Arg37 and the main-chain amide of Ser40, while O3 is hydrogen bonded to Ser40 and a water molecule. This hydrogen-bonding scheme is essentially that observed in src (see, for example, chicken src; PDB code 2ptk), with the exception that in src the contribution of the Ser40 side chain is replaced by the hydroxyl group of a threonine residue which is positioned one amino-acid further in the sequence. Because of the further distance of the phosphotyrosine-binding loop (residues 39–44), this hydrogen-bonding pattern is compromised in the NMR model. It is probable that the reason for this arises from the lower density of NOEs in this region, a difficulty compounded by the fact that many of the potential restraining distances involve unobservable or or very weakly observed labile protons (*e.g.* serine OH, arginine η-NH).

There are some close crystal contacts in the vicinity of the phosphotyrosine. In fact, Arg19, which is hydrogen bonded to the phosphate, also makes a salt bridge to Asp52 in a neighbouring molecule. The phosphotyrosine amino-terminus is wedged between Asp52 and Glu54 in the same adjacent molecule, in a double salt-bridge interaction, and also interacts with the main-chain carbonyl group of Asp52. The peptide ligand N-acetyl group in the NMR structure is not present in the crystal structure; there is, in fact, no room for it in this crystal form. This N-terminal end (pTyr) of the pentapeptide, however, appears to be positionally consistent with other crystal structures; it is the C-terminal end where the larger movements with respect to the NMR structure are observed. The peptide C-terminus makes two interactions with the SH2 domain: there is a hydrogen bond with His94 and a water-mediated interaction with the main-chain carbonyl group of Glu71. Through the same water molecule, there is also an interaction with Tyr107 from another adjacent molecule. It is conceivable that the proximity of adjacent molecules has affected the orientation of the pentapeptide. It is interesting to note, on the other hand, that in a v-src SH2 domain (PDB code 1sha) with the same peptide ligand, the orientation of the peptide is closer to that seen in the p85α C-SH2 crystal structure than in the p85α C-SH2 NMR structure. It is extremely useful to have both NMR and crystal structures of the same protein available in order to prevent over-interpretation of either.

4. Conclusions about the use of NMR trial models

In both these cases, molecular replacement was successful using an ensemble of NMR models from which the more flexible portions (loops and side chains) have been excized. The molecular replacement program *AMoRe* (Navaza, 1994; Collaborative Computational Project, Number 4, 1994) allows one to supply a single PDB file which contains all the atoms of all molecules in the ensemble, thereby enabling, without any particular effort, the use of an ensemble. In our cases, the use of an ensemble provided superior results to the use of a single NMR model and we believe the use of an ensemble is to be recommended. Undeniably, there have been cases of success using a single NMR model or even a minimized mean struc-

ture, but this would not be our first choice. If the intention is to provide maximum opportunity for overlap between the observed and calculated Patterson functions, then use of an ensemble seems reasonable, since for different parts of the molecule different members of the ensemble may prove to be correct. We would discourage use of a minimized mean structure: the mean structure is not an experimental observation itself, it is the average of experimental observations and this carries a risk of being incorrect. The average of two alternate conformations, as an extreme example, is rarely the correct structure. If a mean structure is then minimized to relieve inevitable stereochemical inelegancies obtained from the averaging process, there is added risk, despite inclusion of restraints, of shifting coordinates even further from experimental truth. Nonetheless, unarguable success with the use of average minimized structures has been reported (Chen *et al.*, 2000).

Use of a high-resolution limit that reflects the expected closeness of the trial model to the target structure seems prudent. Values around 4 Å seem reasonable. Careful optimization of the integration radius in the rotation function may help minimize noise from cross-vectors. If an ensemble is used, no special pseudo-temperature factors are required; the spread of possible electron density is inherent in the spread of the ensemble. Cases of success have been reported (Chen *et al.*, 2000) where temperature factors have been devised to account for the observed spread in an ensemble, as suggested by Wilmanns & Nilges (1996) and then applied to, for example, the mean structure; it would seem, however, that applying large temperature factors would be of less use in Patterson overlap than would the presence of all atoms in the ensemble, since large temperature factors would have the effect of flattening electron density, whereas the electron density is maintained in the ensemble at discrete possible locations and thus might contribute more to a Patterson. Use of a single temperature factor chosen sensibly (*e.g.* from the Wilson plot) for all atoms in the ensemble seems a reasonable approach.

One can expect discrimination of the correct solution to be poor and innovative ways of validating the molecular replacement results become valuable. Acceptable crystal packing is a necessary, but far from sufficient, criterion for a correct solution. These are difficult cases (to the extent that alternate phasing methods should be considered if speed is at all of the essence) where inspection of packing and initial electron-density maps may not offer the absolute confirmation required. Clearly, as always, one should aim for the best possible set of diffraction data to maximize chances of success. In the SH2 structure determination described above, the solution was validated by an increasing molecular-replacement signal as the trial model improved, by recognition of an electron-dense marker in the first maps and by consistency between coordinate transformations and cross-crystal rotation functions. In the Im7 structure analysis, discrimination was more difficult, with confidence mainly coming from observing a slightly higher translation-function signal in the correct space group and noting an improvement with rigid-body refinement, as well as the appearance of some side chains in the initial difference electron density. The ultimate confirmation is given by correct refinement: it is debatable whether molecular replacement may be considered successful before the structure is refined. In both cases presented here, refinement posed a significant barrier. At an interim stage, one has oriented and positioned the molecule correctly, but such 'success' in molecular replacement is hollow since nothing new is learned of the target structure. It is not enough therefore in molecular replacement simply to solve the rotation and translation functions: 'solving a structure' should mean obtaining new information, which is revealed by the calculation of electron-density difference maps that, in order to be meaningful, require a reasonable correlation between observed and calculated structure factors, which may require refinement.

What key methodology, then, is responsible for the greater current success with the use of NMR trial models? In both cases reported here, refinement of the structure was initially unsuccessful. When a single model is selected from the ensemble for refinement, inevitably the model becomes less representative of the target structure than the ensemble; it is incomplete and contains many deviations from the final refined X-ray coordinates distributed throughout the structure. This is in general such a poor starting model that it lies outside the radius of convergence for traditional refinement techniques. Only when maximum-likelihood target functions were used was it possible in our examples for the refinement to proceed. The ability of maximum-likelihood refinement to cope well with incomplete and inaccurate starting models is well demonstrated here. It is this recent methodology of incorporation of maximum-likelihood targets in refinement, rather than any new aspect of the molecular replacement, that has allowed success. Thus, the most significant advance for molecular replacement that allows greater use of NMR trial models is not an advance in molecular replacement, but in refinement.

We thank Laurence Pearl and Mark Roe for their efforts with cross-crystal averaging for SH2 and Gerard Kleywegt for useful discussions on Im7 during the 1996 CCP4 molecular replacement workshop in York. We are grateful to the protein science team at AstraZeneca for production of SH2 and to the colicin groups of Richard James and Colin Kleanthous at UEA for the Im7 clone and many discussions.

References

Bacon, D. & Anderson, W. F. (1998). *J. Mol. Graph.* **6**, 219–220.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Booker, G. W., Breeze, A. L., Downing, A. K., Panayotou, G., Gout, I., Waterfield, M. D. & Campbell, I. D. (1992). *Nature (London)*, **358**, 684–687.

Breeze, A. L., Kara, B. V., Barratt, D. G., Anderson, M., Smith, J. C., Luke, R., Best, J. R. & Cartlidge, A. S. (1996). *EMBO J.* **15**, 3579–3589.

Brünger, A. T. (1992). *X-PLOR. Version 3.1. A System for X-ray Crystallography and NMR.* Yale University, Connecticut, USA.
Chen, Y. W. (2001). *Acta Cryst.* D**57**, 1457-1461
Chen, Y. W., Dodson, E. J. & Kleywegt, G. J. (2000). *Structure*, **8**, R213–R220.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Dennis, C. A., Videler, H., Pauptit, R. A., Wallis, R., James, R., Moore, G. R. & Kleanthous, C. (1998). *Biochem. J.* **333**, 183–191.
Esnouf, R. M. (1997). *J. Mol. Graph.* **15**, 133–138.
Hoedemaeker, F. J., Siegal, G., Roe, M., Driscoll, P. C. & Abrahams, J. P. (1999). *J. Mol. Biol.* **292**, 763–770.
James, R., Kleanthous, C. & Moore, G. R. (1996). *Microbiology*, **142**, 1569–1580.
Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Kabsch, W. (1993). *J. Appl. Cryst.* **26**, 795–800.
Kapeller, R. & Cantley, L. C. (1994). *BioEssays*, **16**, 565–576.
Kleanthous, C., Kühlmann, U. C., Pommer, A. J., Ferguson, N., Radford, S. E., Moore, G. R., James, R. & Hemmings, A. M. (1999). *Nature Struct. Biol.* **6**, 243–252.
Ko, T.-P., Liao, C.-C., Ku, W.-Y., Chak, K.-F. & Yuan, H. S. (1999). *Structure*, **7**, 91–102.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Navaza, J. (1994). *Acta Cryst.* D**50**, 157–163.
Nolte, R. T., Eck, M. J., Schlessinger, J., Shoelson, S. E. & Harrison, S. C. (1996). *Nature Struct. Biol.* **3**, 364–374.
Osborne, M. J., Breeze, A., Lian, L.-Y., Reilly, A., James, R., Kleanthous, C. & Moore, G. R. (1996). *Biochemistry*, **35**, 9505–9512.
Pawson, T. & Schlessinger, J. (1993). *Curr. Biol.* **3**, 434–442.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Songyang, Z., Shoelson, S. E., Chausuri, M., Gish, G., Pawson, T., Haser, W., King, F., Roberts, T., Ratnofsky, S., Lechleider, R. J., Neel, B. G., Birge, R. B., Fajardo, J. E., Chou, M. M., Hanafusa, H., Schaffhausen, B. & Cantley, L. S. (1993). *Cell*, **72**, 767–778.
Wallis, R., Leung, K.-Y., Pommer, A. J., Videler, H., Moore, G. R., James, R. & Kleanthous, C. (1995). *Biochemistry*, **34**, 13751–13759.
Wallis, R., Moore, G. R., James, R. & Kleanthous, C. (1995). *Biochemistry*, **34**, 13743–13750.
Wilmanns, M. & Nilges, M. (1996). *Acta Cryst.* D**52**, 973–982.