Structure *14*, 1313–1320, August 2006 ©2006 Elsevier Ltd All rights reserved DOI 10.1016/j.str.2006.06.014

Knowledge-Based Real-Space Explorations for Low-Resolution Structure Determination

Nicholas Furnham,[1,*] Andrew S. Doré,[1,2] Dimitri Y. Chirgadze,[1] Paul I.W. de Bakker,[1,3] Mark A. DePristo,[1,4] and Tom L. Blundell[1]
[1] Sanger Building
Department of Biochemistry
University of Cambridge
Tennis Court Road
Cambridge, CB2 1GA
United Kingdom

Summary

The accurate and effective interpretation of low-resolution data in X-ray crystallography is becoming increasingly important as structural initiatives turn toward large multiprotein complexes. Substantial challenges remain due to the poor information content and ambiguity in the interpretation of electron density maps at low resolution. Here, we describe a semiautomated procedure that employs a restraint-based conformational search algorithm, RAPPER, to produce a starting model for the structure determination of ligase interacting factor 1 in complex with a fragment of DNA ligase IV at low resolution. The combined use of experimental data and a priori knowledge of protein structure enabled us not only to generate an all-atom model but also to reaffirm the inferred sequence registry. This approach provides a means to extract quickly from experimental data useful information that would otherwise be discarded and to take into account the uncertainty in the interpretation—an overriding issue for low-resolution data.

Introduction

Increasingly X-ray crystallographic studies are being directed toward large multiprotein targets with high biological significance. Crystals of such targets often diffract only to low (>3.5 Å) resolutions. Although low-resolution data were usually considered not useful and often abandoned, it is now accepted that they can yield significant insights into biological function (Brunger, 2005).

The difficulty in interpreting low-resolution electron density maps arises from the fact that the number of observations used in their calculation is far smaller than the number of parameters to be defined. In the electron density, this manifests itself as lack of atomicity, often as unresolved peptide groups and tubular density for helices, as well as density accumulating in places other than the main chain, particularly for strands and turns.

*Correspondence: nick@cryst.bioc.cam.ac.uk
[2] Present address: Section of Structural Biology, Institute of Cancer Research, Chester Beatty Laboratories, London, SW3 6JB, UK.
[3] Present address: Department of Molecular Biology, Massachusetts General Hospital, Boston, CPZN-6818, MA 02114.
[4] Present address: Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138.

Some 300 structures have been deposited in the Protein Data Bank (PDB) (Berman et al., 2000) with resolutions of 3.5 Å or worse. An examination of the protocols that have been used in these structures to address the problems of low resolution shows that many involve attempts to improve the phasing, for example by solvent flattening, crystal averaging or symmetry averaging. Most interpret the density by identifying large side chains or posttranslational modifications, often modeling unresolved features or those lacking clear density on the basis of other information or predictions. Restraints and constraints are introduced into refinement in order to reach convergence between the model and the experimental data.

A recent but important example of the determination of protein structures at low resolution is the analysis of unliganded and fully glycosylated SIV gp120 envelope glycoprotein (Chen et al., 2005). Multicrystal averaging was exploited to reduce phase error, and a preliminary model was constructed for 70% of the protein by docking a polyalanine model based on an HIV homolog into the density map. This was followed by iterative cycles of combination of phases from the model and experiment, density modification, data sharpening, model building, and limited rigid-body model refinement. Most protocols have used molecular replacement probes to gain initial phases followed by rigid-body refinement, for example in the structure determination of the acetylcholinesterase tetramer (Bourne et al., 1999) at 4.2 Å resolution. The presence of noncrystallographic symmetry is often exploited to increase the data to parameter ratio and to define phase relationships, for example, in the determination of the truncated human apolipoprotein A-I crystal structure (Borhani et al., 1997) at 4.0 Å resolution. Aromatic residues, selenomethionine residues, and glycosylation sites have been used to identify the register of the sequence in the density. All such approaches rely on initial models being built manually either with fragments or residue by residue.

Several computational methods have been devised to build automatically an initial model based on the density. O (Jones and Kjeldgaard, 1997) and QUANTA (Oldfield, 2001) first trace a Cα skeleton through the density with three-dimensional pattern recognition (Greer, 1974). The Cα points are used to search a database of structures for similar fragments. Once a fragment is found, it is superimposed by least squares fitting onto the Cα skeleton in order to construct the main chain. Sequence decoration is achieved through automated side-chain placement, for example in O, by threading the known sequence onto the structural framework. At each residue in the main-chain trace, a goodness of fit is calculated for each of the 20 possible amino acids (Jones, 2004). Other methods such as RESOLVE (Terwilliger, 2003a, 2003b) do not use skeletonization but employ fragments from a library of refined structures in a hierarchical procedure, first by generating many overlapping fragments and identifying the locations of helical and β strand regions by FFT-based template matching, followed in subsequent stages by extending

and connecting the fragments. Though these methods perform well at high and medium resolutions with minimal user intervention, at lower resolutions, they have proven to be less successful. Problems in skeletonization, such as breaks and branch points, as well as the ambiguity in density features to assign side chains, tend to lead to systematic errors and require a large amount of user intervention to correct.

Here, we describe the application of a conformational search algorithm called RAPPER (DePristo et al., 2003a, 2003b, 2004) to the challenge of interpretation of low resolution structures. RAPPER not only addresses the problem of fitting and modeling unresolved features in unclear density but also significantly contributes to the success of refinement by providing a diverse set of starting conformers that are not necessarily connected by low energy pathways as in the case for molecular dynamics simulations. It also allows the testing of a number of hypotheses based both on the electron density map and on knowledge of homologous structures to generate a model that is consistent with the experimental data. In this way, it can explore ideas about the sequence registry and the location of secondary structure. RAPPER does not address the problem of improving phases. RAPPER can relatively quickly provide an initial all-atom model suitable for refinement. We describe its application to a crystal of ligase interacting factor 1 (Lif1p) (amino acids 1–246) in complex with a fragment of DNA ligase IV (Lig4p) (amino acids 680–944) incorporating the tandem BRCA1 C-terminal (BRCT) of domains from *Saccharomyces cerevisiae*; this complex is involved in the repair of double-stranded DNA breaks in the nonhomologous end-joining DNA repair pathway. The native crystals diffracted to 3.9 Å, and experimental phases were derived from a derivative crystal soaked in 0.5 mM potassium dicyanoaurate (I), diffracting to 4.3 Å resolution. We use this example to discuss the value of RAPPER for interpretation of low-resolution electron density maps.

Results and Discussion

We first tried standard approaches to produce a starting model for refinement, placing ideal secondary structural elements and fragments from previously solved higher-resolution structures of human XRCC4 and homologs of the BRCT domain into the electron density. Attempts were made to refine these chimaeric and partial models by using a variety of strategies including rigid body refinement of the manually placed fragments, refinement by simulated annealing with a variety of restraints, and TLS refinement. However, none resulted in improvement of either R or R free (neither of which dropped below 0.50) nor did they increase the acuity of conformational features or extra density in the electron density maps calculated. Molecular replacement using the models as probes was also unsuccessful both due to the low resolution and to the features of the model probes, in particular the highly repetitive coiled-coil. Automated structure solution methods were unable to generate a starting model.

Given the lack of success of these approaches, we explored the use of RAPPER, which had already proved a powerful tool in the interpretation of crystallographic data at medium and high resolutions (DePristo et al., 2004, 2005). In sampling the energy landscape, RAPPER optimally requires the approximate positions of Cα-atoms as a restraint to guide building, although for short lengths, it can build the polypeptide with only electron density and geometrical constraints. The Cα-atom positions from the manually docked secondary structural fragments were first used as restraints for RAPPER. However, these were often unsatisfactory interpretations of the electron density; in trying to trace through these spatial restraints, and still maintain the model in positive density, RAPPER continually failed to generate a complete model. Skeletonization of the electron density map also proved to be unhelpful in guiding placement of the Cα-atoms due to the low resolution. Large tubes of density, seen most frequently in the coiled-coil region of the Lif1p, tended to become skeletonized as a straight chain, whereas by visual inspection the Cα helical structure could be distinguished. Furthermore, density representing different strands, especially in β sheet regions, had become merged together. To overcome this, Cα-atoms were placed roughly with the Cα baton tool in COOT (Emsley and Cowtan, 2004) guided by the electron density features (Figure 1A) and knowledge of the expected structures from homologs. Though this proved to be a laborious task, it was achieved relatively quickly (a few hours for some 380 residues). This was because there was no requirement for high accuracy placement as RAPPER uses these only as restraints and not actual positions.

The first rounds of rebuilding using the new Cα-atom positional restraints also led to failures resulting from incorrect assumptions about the number of residues present in particular regions, especially at the ends of secondary structural elements. Such errors are easy to introduce at low resolution, as it is very difficult to see how the helices lead into loop regions and to establish the number of residues in a loop. Enforcing secondary structure restraints suggested by inspection of the electron density or from secondary structure predicted from the sequence alignment to structurally related homologs proved a useful restraint in identifying areas of misinterpretation. Failure or difficulty to produce a model consistent with the restraint network was taken to indicate errors in our map interpretation as manifest in the guide positions or secondary structure.

In order to test systematically the effects of errors in the restraints including placement of Cα atom guide points and prediction of secondary structures, two analyses were carried out. First, the sensitivity of the building process to nonsystematic errors in Cα coordinates was assessed by introducing "noise" into the restraints in the form of uniformly distributed random shifts—we refer to them as "errors," although we are not sure the original trace was correct—of varying magnitude into the origins of the Cα restraint spheres from the original trace. A new set of "noisy" restraints was derived for construction of each of the 100 models. This whole operation was repeated with introduced errors in Cα positions of 0.5, 1, 1.5, 2, and 3 Å. This was done for one strand of the Lif1 protomer coiled-coil and the entire Lig4 protomer. The results are shown in Table 1. RAPPER was unable to build models if errors in the Cα atoms positions average greater than 2 Å. If this

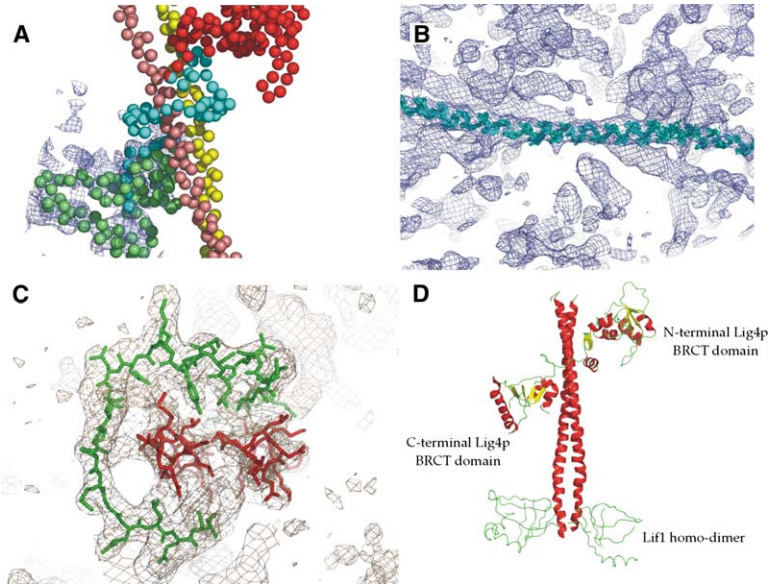

Figure 1. Summary of the Stages in Building an Initial Model and the Final Refined Model

(A) The manually placed Cα atoms shown as space filled atoms in the experimentally derived density at 1σ. For clarity, only part of the electron density map is shown around the C terminus of the Lig4p BRCT domain (green). The linker and N-terminal BRCT domain are shown in cyan and red, respectively. The two coils of the Lif1p are shown in yellow and pink.

(B) An ensemble of ten main chain models (cyan) of one of the Lif1p coils built by using RAPPER with the experimentally determined density at 1.5σ.

(C) An all-atom model looking down the Lif1p coils (red) and part of the Lig4p linker region (green) with the $2F_o–F_c$ density at 1.5σ.

(D) The final refined model shown as a cartoon of the Lif1p, including the two head domain regions and the Lif4p (colored by secondary structure).

condition was met, RAPPER was able to generate all 100 models, with the same rate of success (as measured by the ratio of attempts to build from 100,000 tries). The individual generated conformers have similar rmsds on the average, and the resultant ensemble of solutions is of approximately the same diversity (see Table 1). Thus, RAPPER is unaffected by a reasonable degree of inaccuracy or variation in the Cα guide points.

Table 1. C-Alpha Error Analysis

	Number of Models Built	Number of Times More Than One Pass Was Required	Average Ratio of Attempts to Build a Residue[a]	Ensemble Rmsd[b]	Mean Model Rmsd[b]
Chain A					
Restraint threshold					
1	100	43	0.0454	0.75 (0.08)	0.43
2	100	1	0.1026	1.40 (0.15)	0.68
3	100	1	0.1314	2.04 (0.22)	0.80
4	100	1	0.1684	2.66 (0.29)	0.94
5	100	1	0.2074	3.42 (0.39)	1.24
Noise level					
0.5	100	10	0.0734	1.09 (0.11)	0.63
1.0	100	9	0.0782	1.09 (0.12)	0.65
1.5	100	14	0.0727	1.09 (0.12)	0.66
2.0	100	6	0.0839	1.44 (0.16)	1.10
3.0	0	—	—	—	—
Chain C					
Restraint threshold					
1	100	41	0.0391	0.74 (0.08)	0.46
2	100	1	0.1064	1.39 (0.14)	0.71
3	100	1	0.1685	2.04 (0.21)	0.97
4	100	1	0.2324	2.71 (0.28)	1.27
5	100	1	0.2830	3.39 (0.35)	1.64
Noise level					
0.5	100	10	0.0755	1.07 (0.11)	0.61
1.0	100	14	0.0750	1.07 (0.11)	0.62
1.5	100	37	0.0657	1.07 (0.11)	0.64
2.0	0	—	—	—	—
3.0	0	—	—	—	—

Summary of the effects of changing Cα positions and restraints on the ability of RAPPER to build and the quality of the resulting model for chain A (one of the Lif1 coils) and chain C (the two Lig4 BRCT domains and linker). In each case 100 models were attempted to be built. Two measures of the ease of building are shown: the number of times more than one search pass was required to build a complete model and the average ratio of number of attempts to build a single residue, with the total number of possible attempts is 100,000. Also shown is the ensemble Cα rmsd of the 100 models generated, with the standard deviation shown in brackets. The Cα rmsd of the mean model calculated from the ensemble is also shown.

[a] The ratio is the number of attempts:total number of attempts.

[b] The rmsd is calculated with reference to the deposited refined structure (PDB code: 1Z56).

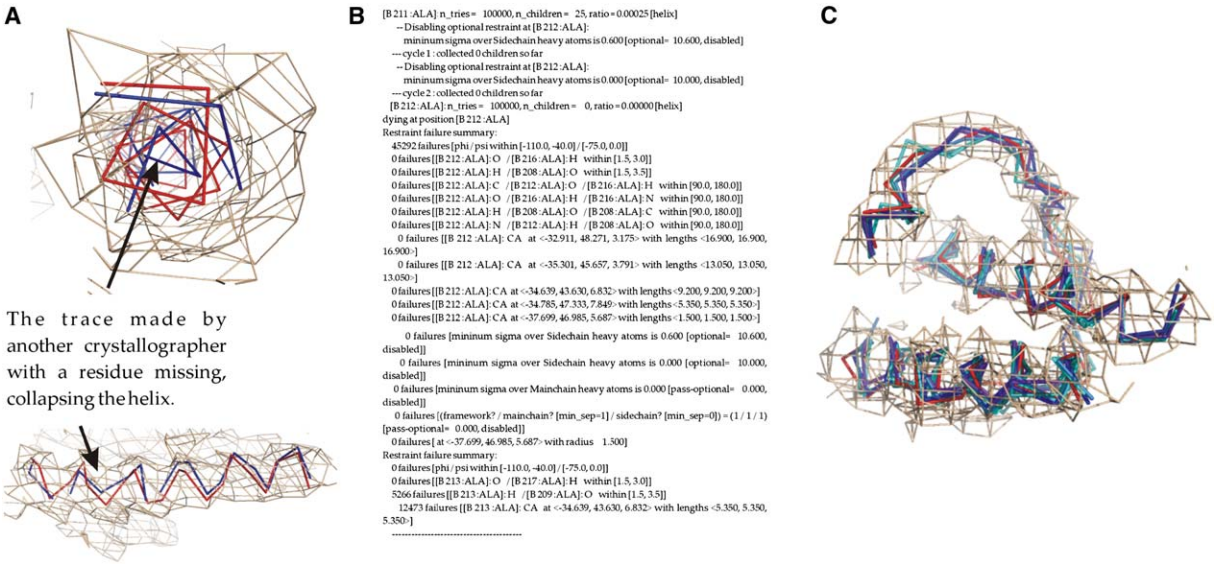

Figure 2. Alternative Tracing by Five Crystallographers

(A) An example of an interpretation in which RAPPER picked up that possibly a residue was misplaced or missing at position 211–212. From the first trace, shown in blue, it can be seen that a residue is missing from the helix, distorting it. The corrected trace is shown in red.

(B) The output by RAPPER surmising the restraint failure from (A). Virtually all the restraints that could not be satisfied were phi/psi indicative of a missing residue.

(C) The results of all five retraces after refinement (shades of blue). For clarity, they are shown as ribbons. The original refined trace is shown in red. The density is the original experimental map at 1.5σ.

A similar analysis was carried out but this time varying the Cα restraint sphere size from 1 Å to 5 Å to assess the required precision of the Cα restraints. Again RAPPER was able to generate all the models requested. Not surprisingly, as the restraints became more stringent, RAPPER found it more difficult to generate models consistent with the restraints. The resultant ensembles are more diverse as the restraints are loosened. Correspondingly, the mean model generated from the ensemble is more distant from the original guide points, though the increase in distance does not directly correspond to the increase restraint sphere size (approximately doubling despite a 5-fold increase in restraint sphere size). This is likely due to the electron density exerting a more dominating restraint than the Cα restraint sphere size.

In order to understand what error individual users may introduce and as an independent test of the robustness of the general method, sections of the original experimental map were retraced by five crystallographers who had not seen the original interpretation or final model. They retraced three sections, one from each of the protomers. The resulting traces were subjected to the RAPPER modeling procedure described above. Several of the traces from the researchers contained "errors" or inconsistencies in the restraints, which prevented RAPPER from building.

A particular problem was found to be the definition of the number of residues in loops between elements of secondary structures as well as missing residues within secondary structure elements. In one such example, a residue was missed in one of the turns of the helix making up one strand of the coiled coil of the Lif1 protomer. Rapper's inability to build resulted in the quick identification of this error. On inspection, it could be quickly seen that if an extra residue were placed in that region, a more helical structure could be produced. After the insertion of the residue, RAPPER successfully and with ease built the entire section (Figures 2A and 2B). In another modeling attempt by one of the researchers, a residue in the linker region of the Lig4 protomer was omitted, and RAPPER was unable to build. Again, when a residue was added at that point, the entire chain could be successfully built. The resultant chain had the same number of residues as determined in the trace made by us originally. An incorrect assessment of the number of residues in a region of the polypeptide chain is usually a fatal error, and thus the position at which RAPPER fails identifies the problematic region.

Once these inconsistencies had been identified with the same criteria for restraint failure as described for the original method and replaced by correct traces, all atom models were generated for each section. The "mean model"—the average of the ten models asked for as in the original model building—for each modeling exercise was then integrated into the prerefined RAPPER model. The mean models were subjected to a single round of rigid body, B-factor by domain, and maximum likelihood refinement in CNS. The resultant models had very similar R (0.43 ± 0.02 with a drop from average of 0.51 to 0.43) and R free (0.51 ± 0.07 with a drop from an average of 0.56 to 0.50) values and Cα rmsd over the retraced sections of less than 2 Å to the original RAPPER prerefined model that had undergone the same round of refinement. The rmsd is within than the optical resolution (the ability to distinguish between two points in space defined as sqrt [2 (sigma_patt^2 + sigma_res^2)] as calculated by SFcheck, CCP4 [Vaguine et al., 1999]) of the map of ~2.5 Å resolution. The refined models are shown in Figure 2C.

Once a self-consistent set of restraints had been formulated, RAPPER quickly built an ensemble of main chain models (see Figure 1B); the run time for a single model of ~380 residues on a Pentium 4 2.3 GHz desktop PC running SuSe Linux 8.2 was 2–3 min. Subsequent to this, sequence register was inferred from the electron density map and assessed by comparing all-atom models with electron density and with Cα and secondary structure restraints. As with the main-chain modeling, errors in sequence registry were identified as sites at which RAPPER had difficulty solving this restraint network. In order to test the impact of changes in sequence registry, a 40 residue section of the Lig4 protomer consisting of two sections of secondary structure flanking a loop region was repeatedly rebuilt with the sequence register offset by one, two, three, and four residues. In each case, RAPPER was asked to generate 100 models for this section. Indications of these errors are much more subtle than for incorrect Cα guide points or errors in the number of residues in a region. RAPPER had a 20%–30% failure rate for small offsets and poor correlation to the density on the models generated for all offsets (see Table 1). In regions with many residues of a similar type, for example those with small side chains such as alanine or serine, the fit to the map is equally good for offset sequence assignments, with very little difference in ability of RAPPER to build. Although RAPPER does generate models that are wrong, it is very good at finding models that can be tested by other means (such as the ability to successfully refine).

These observations demonstrate that RAPPER is useful in modeling to low-resolution data: it enables the testing of weak hypotheses, assumptions, and often speculations about structures suggested by the electron density map. Sequence registry suggested by features in the density as well as information inferred from homologous sequences and structures can be tested. In the case of the BRCT domains, the bulky side chain of a conserved tryptophan supporting one of the β sheets could be distinguished in the map. Similarly the sequence of the Lif1 could be assigned for the linker region between the two BRCT domains that had previously been solved (Sibanda et al., 2001). Efficient exploration of conformational space allows the uncertainty of the dataset to be taken into account; this is important for low-resolution data.

Though marginally more computationally expensive than main-chain building, ten all-atom models were generated, representing a sample of the conformational space consistent with the experimental data. The geometric mean was taken of the ensemble and regularized with Tinker (Ponder and Richards, 1987), so providing a single real-space refined model to take into reciprocal space refinement. The model was first refined with the two helices of Lif1p, the N-terminal BRCT, the linker of the Lig4p, and the C-terminal BRCT domain as rigid body elements. The F_o–F_c difference map showed new features of density not evident in the original electron density. In particular, a section missing between the ends of the first and second helices of the C-terminal BRCT domain was extended. The model also served as a starting point for molecular replacement with PHASER, allowing tentative positioning of the missing head domains of the Lif1p, which had appeared as small, discontinuous sections of density in the original map.

Table 2. Unit Cell and Refinement Statistics

	Derivative	Native
a	250.32	247.62
b	250.32	247.62
c	99.63	98.42
Resolution (Å)	4.2	3.9
Wavelength (Å)	0.975	0.968
Reflections (unique)	142,418 (6,178)	166,810 (16,475)
Completeness[a]	99.9 (100.0)	99.8 (99.6)
R_{SYM}[a]	10.1 (60.0)	10.1 (70.7)
Resolution range (Å)		223.61–3.92
Number of reflections		15,638
Number of non-H protein atoms		4,152
R (%)[a,b]		39.7 (37.7)
R_{free} (%)[a,c]		46.6 (44.3)
Average B factor (Å^2)		110.51
Rms deviation		
Bond lengths (Å)		0.011
Bond angles (°)		1.66

The unit cell and refinement statistics for both the heavy atom derivative and native data set.
[a] Number in parentheses for the outer shell.
[b] R factor = Shkl | | Fo | − | Fc | | / Shkl | Fo |.
[c] R free test set constituted 5.1% (832 reflections) from the working set.

The all-atom model (see Figures 1C and 1D), including most of the Lif1p head domains, was refined with tight restraints by using CNS and REFMAC5 to an R and R free of 0.39 and 0.46, respectively (Table 2). The final refinement statistics compare well with those from other low-resolution structure analyses derived from experimentally derived phase information, where there is no noncrystallographic symmetry and no high similarity high-resolution homologs. They are also consistent with the theoretical limits of data fitting for this observation/parameter ratio and resolution (Tickle et al., 1998, 2000). The R free/R ratio (1.17) is similar to those for the other 1613 structures solved at between 3 Å and 5 Å (Figure 3). However, a large fraction of structures at this resolution (20% at 4 Å or more) has no R or R free statistics, indicating that no refinement was carried out, and a further ~15% have no independent evaluation of the success of the refinement. Independent validation of methods to solve these low-resolution structures is made difficult, however, by the paucity of independently solved high-resolution structures to which they can be compared. Artificial truncations of high-resolution datasets to lower resolution do not provide an adequate benchmark; although the resolution cutoff may be the same, the quality of the diffraction data in the highest resolution ranges is likely to be higher.

Analysis of the final structure (Dore et al., 2006) (PDB code: 1Z56) reveals why other protocols would likely have failed to produce a useful atomic model. The closest structurally resolved homolog of Lif1p (XRCC4) shares less than 20% sequence identity and, unsurprisingly, these two structures differ significantly with respect to the length of the coiled-coil region, the positioning of the two globular N-terminal "head" domains in different planes relative to the Lig4p linker and the

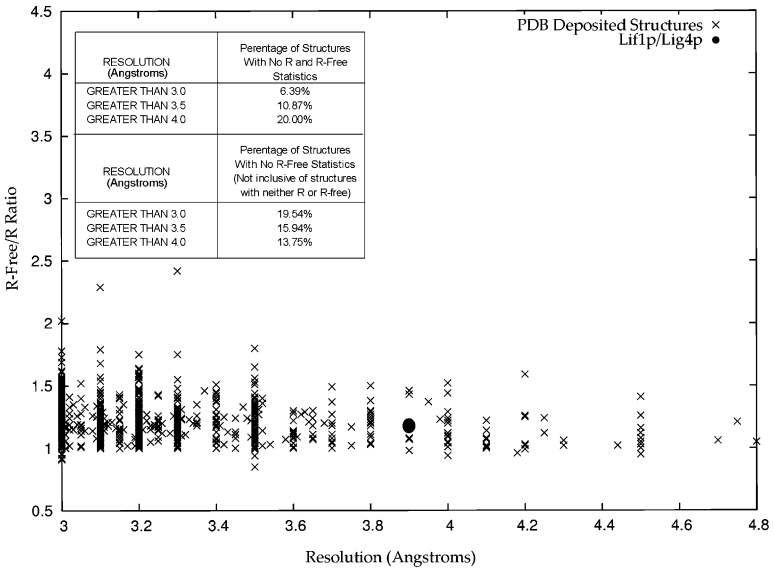

Figure 3. The R Free/R Ratio for all the Deposited X-Ray Structures between 3 and 5 Angstroms in the PDB by Resolution

The inserted table shows the percentage of structures where either an R or R free value was given and the percentage of structures where just an R value was calculated at different resolution cut offs. The 3.9 Å resolution structure of Lif1p-lig4p has an R free/R ratio of 1.17 (indicated by a filled circle). The PDB data were collated from release on 3/22/05.

existence of a kink in the coiled coil in XRCC4. The kink is replaced by a bulge in Lif1p, possibly a consequence of the presence of the longer linking region and the BRCT domains of the Lig4p. It is no wonder that initial attempts at either molecular replacement or docking of ideal secondary structure elements of parts of XRCC4 failed to give a model that could be refined.

The success of the process described here gives confidence that further automation could be achieved. Once the Cα atom guides have been placed, multiple rounds of building, experimenting with different restraint criteria, could be tried and tested. Restraint failure could be automatically tracked, and a decision process similar to that of an interactive user could be used to alter the restraints combinatorially to find a valid solution. Where more than one set of restraints provides a solution, the resultant models could be taken automatically into a refinement strategy and assessed. Reassessment of models that subsequently refined poorly could feed back to inform further attempts at model building. An example of where this would be of most benefit is in sequence assignment. In regions with high prevalence of residues of a similar type, the fit to the map is equally good with very little difference in ability for RAPPER to build. Thus, if the correct number of residues has been placed in that region, but the entire sequence has slipped, then it is more than likely that RAPPER will generate a model, but its refinement will be compromised. For all these combinations of possible restraints being tested in a combinatorial manor, a computer is a much better tool for tracking than a human user.

An efficient protein conformational search engine that takes into account the experimental electron density, protein stereochemistry, and spatial considerations enabled us to produce quickly an initial model that is consistent with both the experimental data and the a priori information about protein structure. The use of these additional restraints allowed unambiguous determination of the sequence register from a few seeding points. The resultant model reached a point that it could be taken into further refinement. Our approach to treating low-resolution data already enables us to extract biologically useful information from experimental data (Dore et al., 2006) that might otherwise be unused or even discarded.

Experimental Procedures

Data were collected from Lig4p:His6-Lif1p cocrystals (space group $P6_422$ a = b = 247.62 and c = 98.42 Å). A single crystal heavy atom derivative diffracted to 4.3 Å resolution at a wavelength of 0.975 Å, while the native crystals diffracted to 3.9 Å at a wavelength of 0.968 Å (Dore et al., 2006). Experimental data of native and derivative crystals were processed with the program HKL (Otwinowski and Minor, 1997). Four gold sites were detected with the Shake'N'Bake program (Weeks and Miller, 1999), with phases determined by SHARP (Bricogne et al., 2003) and the resultant density modified by SOLOMON (Abrahams and Leslie, 1996). Cα positions were loosely built into the subsequent electron density map manually by using the Cα-baton tool within COOT (Emsley and Cowtan, 2004).

RAPPER was then used to build ten polyalanine models with the manually positioned Cα-atom points as guides. Two ensembles of models were built with 1 Å and 2 Å distance restraints from the Cα-atom points to limit the space in which RAPPER could search for a valid solution. Restraints in the model building included the electron density, as well as those implemented in the RAPPER algorithm deriving from residue-specific φ/ψ propensity tables, ideal geometry, and excluded volume restraints (DePristo et al., 2004). Secondary structure restraints of restricted φ/ψ angles and hydrogen-bonding distance restraints were also enforced. A schematic summary is given in Figure 4. The existence of secondary structure was inferred from visual inspection of the density and from the sequence alignment with homologs of known structure. Errors in the number of residues and placement were identified by monitoring the RAPPER build process.

The failure of RAPPER to build was interpreted by assessing which of the restraints RAPPER was unable to satisfy. A summary of the restraint failure is given (see Figure 2B) at the last residue attempted to be built. Failure to build through inability to satisfy phi/psi restraints was indicative of an incorrect number of residues being assigned to a region. Residues were added or deleted (one at a time) and RAPPER rerun until a set of self-consistent restraints was achieved. Where the failure was primarily due to not being able to satisfy the electron density, then the position of the Cα guide point was reassessed and moved into a region of stronger density. RAPPER also removes or increases the size of restraints in an attempt to continue building. Areas requiring a reassessment of the modeling are identified by monitoring where and what the restraints are and which are being removed or changed. Often the electron density restraints are loosened by reducing the sigma level cutoff.

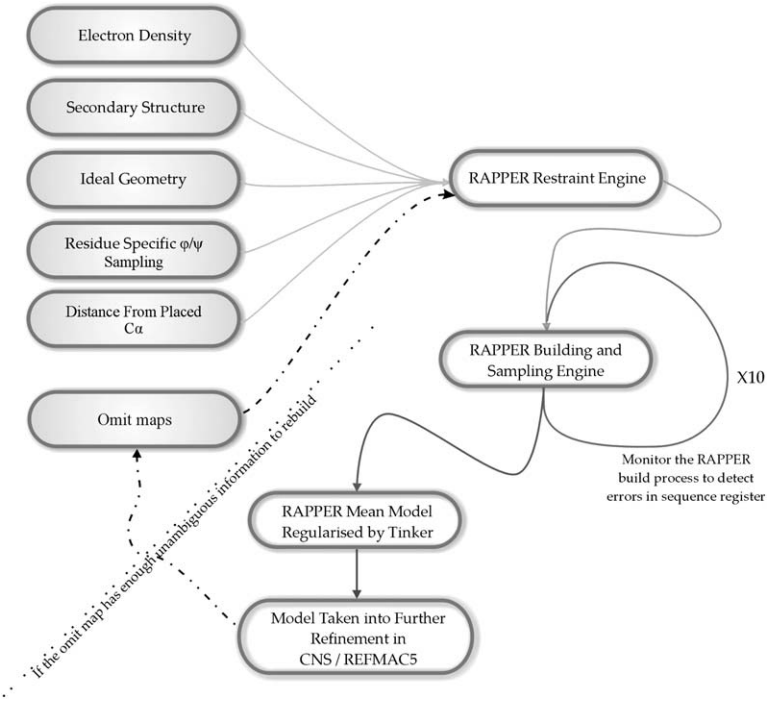

Figure 4. Schematic of the RAPPER Build and Refinement Process

The types of restraints used in the RAPPER restraint engine are indicated.

Sequence was assigned on the basis of three anchor points with distinguishing features in the density. The first anchor point comprised the highly conserved interaction site residues between the Lif1p/Lig4p determined by homology to the high-resolution human homolog XRCC4 (Junop et al., 2000; Sibanda et al., 2001). The other two points were in the BRCT domains of the Lig4p. One was located in a helix that has a highly conserved tryptophan residue supporting the β-sheet region above, the large aliphatic side chain of which can be discerned as a large bulge in the density. The other was a histidine/glycine/glycine right-angle turn, which was also visible in the density. We produced ten all-atom models with RAPPER. Assumptions in sequence registry were assessed by monitoring the RAPPER build process. Where there were errors in building the sequence into density, especially where residues had been inserted or deleted relative to the correct alignment, RAPPER often struggled to find a solution consistent with the Cα, residue type, and density restraints. These revealed areas were inaccuracies may have occurred. In a similar manner to the assessment of restraint failure for the number of residues and placement of the Cα guide points, failure to build after assigning sequence was interpreted by assessing the type of restraint that could not be satisfied. Failure to build due to inability to satisfy electron density restraints or where electron density restraints were disabled or loosened resulted in iteratively shifting the sequence the of entire section being built by one up or down, until a self-consistent set of restraints could be generated. This was done in conjunction with the sequence alignment of homologous structures and sequences. Where more than one conformer could be generated, the quality of the fit was assessed by visual inspection, and the score of electron density fit as implemented in COOT. The mean structure of the all-atom model ensemble was calculated by RAPPER and regularized by TINKER (Ponder and Richards, 1987). This final model was then taken into further refinement.

No models could be produced for two domains of 150 residues located at the N terminus of the Lif1p due to the poor quality and connectivity of the density. The lack of density in this region is probably due to the inherent dynamic disordering of these globular domains as they lie in a large solvent channel down the 3-fold screw axis of the crystal lattice and are probably only making minimal contacts between each other. The model of the main structure was used as a molecular replacement probe in PHASER (Storoni et al., 2004), along with two probes of the missing domains derived from the XRCC4 homolog consisting of just the secondary structural elements. One domain was able to be placed by PHASER, while the other was placed manually by assuming noncrystallographic 2-fold symmetry along the Lif1p coiled-coil axis. A round of rigid body refinement on the manually placed domain was followed by domain B-factor refinement and one round of 100 cycles of maximum likelihood refinement with the CNS (Brunger et al., 1998) package of the whole structure. This was followed by a final round of highly restrained maximum likelihood refinement with REFMAC5 (Steiner et al., 2003). The final model was assessed by PROCHECK (Laskowski et al., 1993) and by the R and R free refinement statistics calculated by REFMAC5.

To address issues of the effect of initial interpretation of the map and placement of the Cα atom guide points, three sections of the map were independently interpreted by five crystallographers. The three sections comprised of two sections of helix from each side of the Lif1 protomer coiled-coil and a section of the linker region of the Lig4 protomer. The original experimental map was used, and information of the existing homologous structures was also given. Once the sections had been traced, they were then subjected to the same RAPPER building strategy as described above of first generating an ensemble of ten polyalanine models, assigning sequence and building an ensemble of ten all-atom models. At each stage, if RAPPER was unsuccessful in generating a model consistent with the restraints, the model was reevaluated and altered with the same criteria for restraint failure interpretation as described earlier. The mean structure of the all-atom models was calculated and geometrized with TINKER. This final model was then integrated into with the rest of the structure as generated by RAPPER after in the initial tracing. These five models were then taken into a round rigid body, B-factor by domain, and maximum likelihood refinement with CNS. The model generated by the initial tracing was also taken through the same refinement procedure for comparison purposes.

In order to assess the effects of errors in the placement of the Cα guide points, uniformly distributed, random shifts/errors of varying magnitude were introduced into the origins of the Cα restraint spheres. A new set of "noisy" restraints was derived for each pass. RAPPER was then used to attempt to build 100 models. This was repeated introducing increasingly larger magnitudes of noise of 1.0, 1.5, 2.0, and 3.0 Å. In a similar manner, the restraints enforced on the Cα positions were systematically changed by with increasing radii from 1 Å to 5 Å, incrementing by 1 Å. Again, attempts were made to generate 100 models. The effects of varying Cα position and other restraints were conducted on one of the coils of the Lif1 protomer and the entire Lig4 protomer. Sequence assignment was also tested

by systematically altering the sequence register in a 40 residue region of the Lig4 protomer, spanning an extended loop region between two elements of secondary structure. The number of residues was not altered, but the sequence assignment was slipped by one, two, three, and four places. For each new assignment, RAPPER attempted to build 100 models.

Acknowledgments

We thank the following for retracing the structure: Stefania Ragone, Nicholas Harmer, J. Venkatesh Pratap, Anjum Karmali, and Rick Smith. N.F. and A.S.D. were supported by Biological and Biological Science Research Council studentships. P.I.W.d.B. was supported the Cambridge European Trust, the Isaac Newton Trust, and the Biological and Biological Science Research Council. M.A.D. was supported by the Marshall Aid Commemoration Commission, U.S. National Science Foundation, and the Cambridge Overseas Trust and is now a Damon Runyon Fellow supported by the Damon Runyon Cancer Research Foundation (DRG-1861-05). D.Y.C. was supported by a grant from the Wellcome Trust.

Received: March 7, 2006
Revised: June 12, 2006
Accepted: June 16, 2006
Published: August 15, 2006

References

Abrahams, J.P., and Leslie, A.G.W. (1996). Methods used in the structure determination of bovine mitochondrial F-1 ATPase. Acta Crystallogr. D Biol. Crystallogr. *52*, 30–42.

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

Borhani, D.W., Rogers, D.P., Engler, J.A., and Brouillette, C.G. (1997). Crystal structure of truncated human apolipoprotein A-I suggests a lipid-bound conformation. Proc. Natl. Acad. Sci. USA *94*, 12291–12296.

Bourne, Y., Grassi, J., Bougis, P.E., and Marchot, P. (1999). Conformational flexibility of the acetylcholinesterase tetramer suggested by X-ray crystallography. J. Biol. Chem. *274*, 30370–30376.

Bricogne, G., Vonrhein, C., Flensburg, C., Schiltz, M., and Paciorek, W. (2003). Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0. Acta Crystallogr. D Biol. Crystallogr. *59*, 2023–2030.

Brunger, A.T. (2005). Low-resolution crystallography is coming of age. Structure *13*, 171–172.

Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., et al. (1998). Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr. D Biol. Crystallogr. *54*, 905–921.

Chen, B., Vogan, E.M., Gong, H., Skehel, J.J., Wiley, D.C., and Harrison, S.C. (2005). Determining the structure of an unliganded and fully glycosylated SIV gp120 envelope glycoprotein. Structure *13*, 197–211.

DePristo, M.A., de Bakker, P.I., Lovell, S.C., and Blundell, T.L. (2003a). Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles. Proteins *51*, 41–55.

DePristo, M.A., De Bakker, P.I., Shetty, R.P., and Blundell, T.L. (2003b). Discrete restraint-based protein modeling and the Calpha-trace problem. Protein Sci. *12*, 2032–2046.

DePristo, M.A., de Bakker, P.I., and Blundell, T.L. (2004). Heterogeneity and inaccuracy in protein structures solved by x-ray crystallography. Structure *12*, 831–838.

DePristo, M.A., de Bakker, P.I., and Blundell, T.L. (2005). Crystallographic refinement by knowledge-based exploration of complex energy landscapes. Structure *13*, 1311–1319.

Dore, A.S., Furnham, N., Davies, O.R., Sibanda, B.L., Chirgadze, D.Y., Jackson, S.P., Pellegrini, L., and Blundell, T.L. (2006). Structure of an Xrcc4-DNA ligase IV yeast ortholog complex reveals a novel BRCT interaction mode. DNA Repair (Amst.) *5*, 362–368.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. *60*, 2126–2132.

Greer, J. (1974). Three-dimensional pattern recognition: an approach to automated interpretation of electron density maps of proteins. J. Mol. Biol. *82*, 279–301.

Jones, T.A. (2004). Interactive electron-density map interpretation: from INTER to O. Acta Crystallogr. D Biol. Crystallogr. *60*, 2115–2125.

Jones, T.A., and Kjeldgaard, M. (1997). Electron-density map interpretation. Macromol. Crystallogr. B *277*, 173–208.

Junop, M.S., Modesti, M., Guarne, A., Ghirlando, R., Gellert, M., and Yang, W. (2000). Crystal structure of the Xrcc4 DNA repair protein and implications for end joining. EMBO J. *19*, 5962–5970.

Laskowski, R.A., Macarthur, M.W., Moss, D.S., and Thornton, J.M. (1993). Procheck—a program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. *26*, 283–291.

Oldfield, T.J. (2001). A number of real-space torsion-angle refinement techniques for proteins, nucleic acids, ligands and solvent. Acta Crystallogr. D Biol. Crystallogr. *57*, 82–94.

Otwinowski, Z., and Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. Macromol. Crystallogr. A *276*, 307–326.

Ponder, J.W., and Richards, F.M. (1987). An efficient newton-like method for molecular mechanics energy minimization of large molecules. J. Comput. Chem. *8*, 1016–1024.

Sibanda, B.L., Critchlow, S.E., Begun, J., Pei, X.Y., Jackson, S.P., Blundell, T.L., and Pellegrini, L. (2001). Crystal structure of an Xrcc4-DNA ligase IV complex. Nat. Struct. Biol. *8*, 1015–1019.

Steiner, R.A., Lebedev, A.A., and Murshudov, G.N. (2003). Fisher's information in maximum-likelihood macromolecular crystallographic refinement. Acta Crystallogr. D Biol. Crystallogr. *59*, 2114–2124.

Storoni, L.C., McCoy, A.J., and Read, R.J. (2004). Likelihood-enhanced fast rotation functions. Acta Crystallogr. D Biol. Crystallogr. *60*, 432–438.

Terwilliger, T.C. (2003a). Automated main-chain model building by template matching and iterative fragment extension. Acta Crystallogr. D Biol. Crystallogr. *59*, 38–44.

Terwilliger, T.C. (2003b). Automated side-chain model building and sequence assignment by template matching. Acta Crystallogr. D Biol. Crystallogr. *59*, 45–49.

Tickle, I.J., Laskowski, R.A., and Moss, D.S. (1998). R-free and the R-free ratio. I. Derivation of expected values of cross-validation residuals used in macromolecular least-squares refinement. Acta Crystallogr. D Biol. Crystallogr. *54*, 547–557.

Tickle, I.J., Laskowski, R.A., and Moss, D.S. (2000). R-free and the R-free ratio. II. Calculation of the expected values and variances of cross-validation statistics in macromolecular least-squares refinement. Acta Crystallogr. D Biol. Crystallogr. *56*, 442–450.

Vaguine, A.A., Richelle, J., and Wodak, S.J. (1999). SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. Acta Crystallogr. D Biol. Crystallogr. *55*, 191–205.

Weeks, C.M., and Miller, R. (1999). The design and implementation of SnB version 2.0. J. Appl. Crystallogr. *32*, 120–124.