Acta Cryst. (1996). D**52**, 129–142

Structural Refinement of the DNA-Containing Capsid of Canine Parvovirus using *RSRef*, a Resolution-Dependent Stereochemically Restrained Real-Space Refinement Method

MICHAEL S. CHAPMAN[a,b] AND MICHAEL G. ROSSMANN[a]*

[a]*Department of Biological Sciences, Purdue University, West Lafayette, IN 47907-1392, USA, and* [b]*Department of Chemistry and Institute of Molecular Biophysics, Florida State University, Tallahassee, FL 32306, USA*

(*Received 23 February 1995; accepted 1 June 1995*)

Abstract

The canine parvovirus structure (CPV) [Tsao, Chapman, Agbandje, Keller, Smith, Wu, Luo, Smith, Rossmann, Compans & Parrish (1991). *Science*, **251**, 1456–1464] has been refined by a real-space refinement procedure [Chapman (1994). *Acta Cryst.* A**51**, 69–80]. The fit of an atomic model to electron density was optimized while taking into account the resolution limit of the data and the stereochemistry of the structure. The refined model had a reasonable free *R* factor [Brünger (1992). *Nature (London)*, **355**, 472–475] of 0.29. The method is particularly fast and convenient when only a small fraction of the crystallographic asymmetric unit needs to be refined, as is the case when there is high non-crystallographic redundancy. Cycles of refinement for virus capsids were completed in about 1/50th of the time required for equivalent reciprocal-space procedures.

1. Introduction

1.1. *Objectives*

The refinement of virus structures by methods such as *PROLSQ* (Hendrickson, 1985) or *X-PLOR* (Brünger, Kuriyan & Karplus, 1987) is an intensely time consuming process. Many structures remain unrefined, especially viruses modified by mutations, ligands or pH, where the benefits of refinement (often with respect to partial data sets) do not justify the effort. Chapman (1994) described a real-space approach to refinement that is about 50 times faster in the presence of 60-fold non-crystallographic redundancy. We present here a description of an application of this procedure to the refinement of the monoclinic canine parvovirus structure (Tsao *et al.*, 1991).

1.2. *Refinement method*

Instead of optimizing the agreement of structure amplitudes calculated from an atomic model with those measured experimentally, the real-space method described by Chapman (1994), like that of Diamond (1971), and Jones & Liljas (1984) minimizes a residual (R_ρ) of the difference between experimental electron density, ρ_o, and the density, ρ_c, calculated with respect to the parameters (P) that define an atomic model, where,

$$R_\rho(P) = \sum_{xyz}[S\rho_o(xyz) + k - \rho_c(xyz)]^2, \qquad (1)$$

and S and k are refinable scaling parameters. The summation is over all xyz grid points near atoms. Unlike other real-space refinement methods, ρ_c is calculated by Fourier transformation of resolution-truncated scattering factors (Chapman, 1994), so that medium-resolution maps can be modelled precisely. This makes it possible to refine individual atoms with stereochemical restraints, rather than constrained groups of atoms. R_ρ is then combined with $R_{\rm geom}$, the residual of the difference between current and ideal model geometry, for simultaneous refinement of stereochemistry and fit to the map.

Atomic positions are refined with respect to the current best electron-density map in real-space refinements. Real-space refinement is, therefore, implicitly dependent upon the phases. Thus, for many applications in protein crystallography, real-space refinement is at a disadvantage relative to reciprocal-space refinement, because experimental phases are often imprecise, and phases calculated from a model can be biased. However, when the non-crystallographic redundancy is high, as in the structure determination of viruses, the phases are likely to be especially accurate. Arnold & Rossmann (1988) showed that the use of phases as targets for refinement, in addition to structure amplitudes, helped in the refinement of rhinovirus 14. Viral data sets contain hundreds of thousands of structure amplitudes and it is common (Silva & Rossmann, 1985; Arnold & Rossmann, 1988) to reduce computation for reciprocal-space refinement by using a series of subsets of the data. In contrast, all of the structure amplitudes and phases are used to compute the electron-density maps into which models are refined in real-space methods.

1.3. *Canine parvovirus*

Parvoviruses are small and contain single-stranded (ss) DNA genomes of about 5 kb (Berns, 1990), surrounded by a protein shell containing 60 capsid proteins related by icosahedral symmetry (Tsao *et al.*, 1991). They generally replicate in highly proliferating cells. The human parvovirus, B19, is associated with aplastic crisis, childhood fifth disease and spontaneous abortions (Anderson & Török, 1989; Kurtzman *et al.*, 1989). The canine version is often fatal to young dogs (Studdard, 1990). The CPV structure (Tsao *et al.*, 1991) showed that the capsid protein has a β-barrel motif similar to that found in many RNA viruses, as well as double-stranded DNA viruses, but with large loop insertions that decorate the viral surface and often form antigenic epitopes. The CPV structure also showed 11 nucleotide fragments of the ssDNA on the inside surface of the capsid, bound to each of the icosahedrally related capsid proteins.

1.4. *Nucleic acid binding site*

The nucleic acid is generally not seen in virus structures, presumably because it mostly lacks the icosahedral symmetry that is used for structure determination. There have been a few examples where fragments of nucleic acid have been seen: helical tobacco mosaic virus (Stubbs, 1989), bacteriophages φX174 (McKenna, Xia, Willingmann, Ilag & Rossmann, 1992) and MS2 (Välegard, Murray, Stockley, Stonehouse & Liljas, 1994), bean pod mottle virus (BPMV) (Chen *et al.*, 1989), Flock House virus (Fisher & Johnson, 1993) and satellite tobacco mosaic virus (Larson *et al.*, 1993). In CPV, 11 ordered nucleotides per icosahedral asymmetric unit, or 660 per virion, were seen (Tsao *et al.*, 1991). There is one DNA binding site for each of the 60 symmetry-related proteins of the capsid. The strength of the electron density suggests that most of the sites are occupied. However, the genomic sequence of CPV does not contain any sequences that repeat 60 times. Therefore, the nucleic acid that is seen is the superposition or average of different sequences. At some of the positions, the observed electron density or the stereochemical environment shows that there is a preference for a specific type of base. The preferred base types were used in model building and resemble a consensus binding sequence.

1.5. *Initial structure determination*

The initial structure determination of monoclinic crystals of DNA-containing CPV particles was reported by Tsao *et al.* (1992), but relevant aspects will be summarized here. Data were collected on film at synchrotrons using 0.4° oscillations, and were processed and postrefined (Rossmann, 1985) to yield $R_{\rm merge}$ of 12.2% for 583747 unique reflections derived from 970770 observations greater than $2\sigma(I)$. The 2σ data set was over 80% complete between ∞ and 5 Å resolution, but only $\sim$ 40% complete between 3.0 and 3.5 Å resolution and was increasingly sparse to the limit of processing, namely 2.8 Å resolution. The particle orientation was determined by a self-rotation function (Rossmann & Blow, 1962). The r.m.s. error between the rotation-function peaks and corresponding directions of a 'standard' icosahedron had an angular error of 0.05°, equivalent to a 0.12 Å positional error on the surface of the virus. The position of the particle was refined by several methods as the phases were extended to higher resolution, with the final position being the one that gave the lowest deviation between electron densities related by icosahedral symmetry. Low-resolution phases were initially determined by fitting spherical shell models to the diffraction data (Chapman, Tsao & Rossmann, 1992; Tsao, Chapman & Rossmann, 1992). With hindsight, these phases might have been good enough for extension to high resolution, but they were used merely to solve the position of a single isomorphous (SIR) derivative at 9 Å resolution. The SIR phases were refined and extended by icosahedral 60-fold symmetry averaging (Rossmann, 1990) to 3.25 Å resolution. Attempts to extend the phases to 3.0 Å resolution were abandoned after obtaining low correlation coefficients and maps that were indistinguishable from those calculated at 3.25 Å. A model built using *FRODO* (Jones, 1978) yielded an unrefined conventional R factor of 35% for data between 5 and 3.25 Å resolution and was the basis of the initial report of the structure (Tsao *et al.*, 1991).

The first model to be refined was that from a different, tetragonal crystal form containing empty particles (Wu, Keller & Rossmann, 1993). Its structure was determined by molecular replacement, using the unrefined monoclinic full capsid structure (whose refinement is described here). Phases for the empty capsid structure were refined to 3.5 Å resolution using 30-fold non-crystallographic redundancy. The tetragonal structure was refined with *X-PLOR* (Brünger *et al.*, 1987) and *PROLSQ* (Hendrickson, 1985) modified to minimize,

$$R = \sum_h \{(1 - w_h)(|F_h^{\rm obs}| - |F_h^{\rm calc}|)^2 + (w_h)[(A_h^{\rm mr} - A_h^{\rm calc})^2 + (B_h^{\rm mr} - B_h^{\rm calc})^2]\} + R_{\rm geom}, \qquad (2)$$

where (A_h, B_h) are the real and imaginary components of the structure factor, F_h, of reflection h. The mr superscripts refer to structure factors with phases determined by (molecular replacement) real-space electron-density averaging, and calc denotes atomic structure-factor calculations. The weights, w_h, are the figures of merit which vary from 0.0 to 1.0. If all w_h were zero, this would correspond to conventional *PROLSQ*. The final refinement included 84 water molecules and isotropic temperature factors, giving an R factor of 0.21 with the geometrical statistics listed in Table 1.

2. Methods

Real-space refinement was implemented as an additional module (*RSRef*; Chapman, 1994) for the refinement package *TNT* (Tronrud, Ten Eyck & Matthews, 1987). The real-space refinement module substituted for *TNT*'s *RFactor* program that is normally used to calculate the structure-factor terms. First- and second-order derivatives

Table 1. *Selected stereochemical statistics for the full and empty capsids at the completion of refinement*

All statistics were calculated by *TNT*'s *Geometry* (Tronrud *et al.*, 1987). The deviation of torsion angles is higher than often reported because the main chain φ, ψ angles were included.

R.m.s. variation from ideal	Empty capsid	DNA-containing capsid
Bond lengths (Å)	0.018	0.023
Bond angles (°)	2.7	4.4
Torsion angles (°)	15.3	19.6
Close contacts (Å)	0.19	0.29
No. of poor contacts between protomers	157	94

of the electron-density residual (1) with respect to the atomic parameters were combined in *TNT*'s *Shift* routine with derivatives of the stereochemical residual (from *TNT*'s *Geometry* routine) to calculate a shift vector for the atomic parameters. Although it is necessary to refine only the atoms in one non-crystallographic asymmetric unit, atoms from neighboring asymmetric units are required to define inter-protomer contacts and to calculate the total electron density for pixels where atoms from neighboring protomers make overlapping contributions at low resolution. The individual steps of a cycle are shown in Table 2.

Following the nomenclature of Tronrud *et al.* (1987), a 'long' cycle contains 'short' cycles repeated until convergence. At the start of each long cycle a new direction is calculated for the shift vector using the derivatives. As the refinement is non-linear, the optimizations of different parameters are interdependent. Several short cycles are used to optimize the magnitude of the shift in the direction determined at the start of the long cycle. The short cycles are quicker, because only the residuals, and not the derivatives, need to be calculated. Several long cycles were required to achieve convergence with a particular set of weights. Generally, these were run without intervention and are collectively known here as a 'batch'. Continued improvement was often possible with successive batches of refinement, between which the weights were adjusted. Here, these are collectively called a 'round', after which further improvement was possible only with remodeling. Four rounds of real-space refinement, alternated with interactive model building, were used to refine the structure of DNA-containing CPV as summarized in Fig. 1 and as detailed in the following sections.

The progress of refinement was monitored through the calculation of R factors against small randomly selected subsets of the data. These subsets were small to enable quick, repeated R-factor calculation for the ~300 000 atoms of the CPV asymmetric unit. Tests with different subsets showed that R factors could be calculated with a standard error of 0.01 (on an absolute scale) with subsets of 500 reflections. To facilitate comparisons, the R factors used to monitor refinement (Tables 3 to 9) were calculated with subsets of ~500 or ~1000 reflections with the same selection criteria (between ∞ and 3.25 Å resolution, with $F > 5\sigma_F$). A $5\sigma_F$ cut off was used to slightly emphasize the better measurements in the monitoring process. Conventional R factors, $R^{\rm conv}$, for rounds 1 and 2 were always calculated using the same subset, as were the free R factors of rounds 3 and 4 with another subset. Comparisons of R factors calculated against the same subset of data are not limited by the absolute precision of the R factors. The free R factors for the final model (Table 10) were calculated to a precision of about ±0.0025 (*i.e.* ±0.25%) by using ~7000 randomly selected reflections, without any F/σ_F selection.

Table 2. *Steps within one cycle of real-space refinement*

The table lists programs from *TNT* (Tronrud *et al.*, 1987) and the *RSRef* package (Chapman, 1994). Each 'long' cycle contains several repeats of a 'short' cycle. Within the first of the 'short' cycles, derivatives are needed so that a new direction for the shift vector can be calculated (Tronrud *et al.*, 1987). Subsequent 'short' cycles are quicker, because residuals, but not derivatives, are calculated to optimize the length of the vector whose direction is not reoptimized until the next 'long' cycle.

Program	Package	Function
(1) *Expcoord*	*RSRef*	Expand the coordinates to include symmetry equivalent neighboring residues (crystallographic and non-crystallographic symmetry) with which refining atoms may interact stereochemically or have overlapping electron density.
(2) *Rsref*	*RSRef*	Calculate residual (all cycles) and derivatives (1st 'short' cycle of each 'long' cycle) of fit to electron density for each atom of the non-crystallographic asymmetric unit.
(3) *Geometry*	*TNT*	Calculate stereochemical residual and derivatives (1st 'short' cycle only) for symmetry-expanded coordinates.
(4) *Geometry*	*TNT*	Repeated for the symmetry-expanded coordinates that were **not** in the original non-crystallographic asymmetric unit.
(5) *RmDeriv*	*RSRef*	Subtract the terms generated in step 4 from those of step 3, to generate a unique set of stereochemical terms for one non-crystallographic equivalent that includes interactions with other symmetry equivalents.
(6) *Shift*	*TNT*	Calculate atomic shifts for the non-crystallographic asymmetric unit.
(7) *ReExpand*	*RERef*	Apply corresponding shifts to symmetry equivalents.
(8)		Return to step 2, beginning a new 'long' cycle whenever 'short' cycles have converged.

2.1. *Initial real-space refinement*

Prior to automatic refinement, parts of the model of Tsao *et al.* (1991) with poor stereochemistry or fit to the electron density were rebuilt using the interactive graphics program, *O* (Jones, Zou, Cowan & Kjeldgaard, 1991). The map that was used for this rebuilding and the subsequent refinement was averaged according to the 60-

fold non-crystallographic symmetry from a map that had been calculated with structure amplitudes that were weighted (as previously described, Tsao, Chapman, Wu *et al.*, 1992) according to the agreement between observed amplitudes and those back-transformed from the previous cycle of electron-density averaging. The first round of refinement, being the first application of the real-space refinement procedure, was somewhat experimental (Table 3). Several different weighting schemes were tried. Following common practice with *PROLSQ*, the geometrical restraints were alternately loosened then tightened in the hope of increasing the convergence radius, should any regions of the model be held in incorrect conformation by tight stereochemical restraints. In batch 2 the r.m.s. geometrical errors were allowed to rise to 3.0 Å for bond lengths and 21° for bond angles, before being more tightly restrained. In batch 3, the errors rose to 0.06 Å and 8°, respectively. The usefulness of alternate stereochemical relaxation and tightening was tested by repeating batch 3 (as batch 3*b*) with weights chosen to refine steadily towards target values. The resulting statistics are similar, but fewer cycles were required in batch 3*b*. The improvement in torsion-angle geometry in batch 4 was associated with the realization that by using *TNT* defaults, they had previously been given zero weight. The product of the four batches that constituted round 1 was named RS1.

Model RS1 was selectively remodeled, paying closest attention to regions that were poor according to several criteria: firstly, the residues listed by *TNT*'s *Geometry* (Tronrud *et al.*, 1987) as containing the worst 20 deviations from ideality for bond length, bond angle or

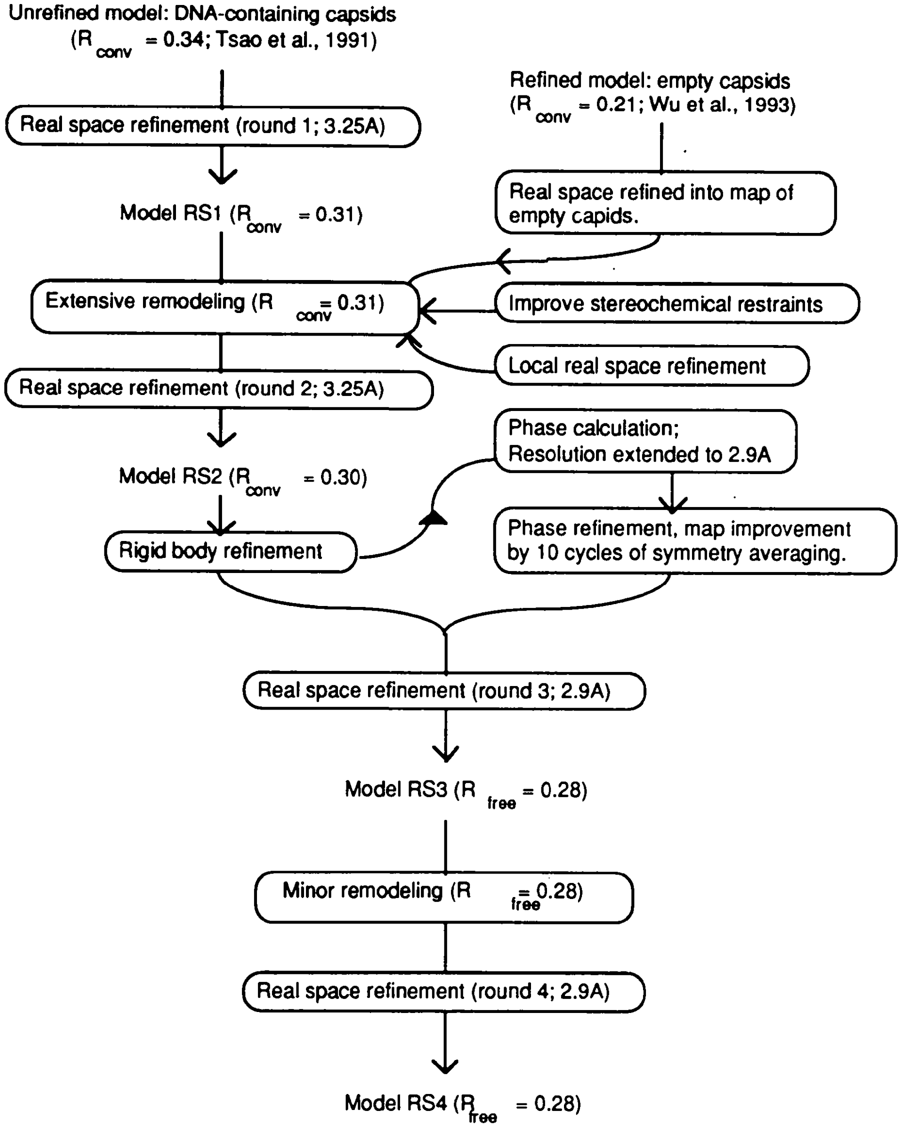
Unrefined model: DNA-containing capsids
(R conv = 0.34; Tsao et al., 1991)
Real space refinement (round 1; 3.25A)
Model RS1 (R conv = 0.31)
Extensive remodeling (R conv = 0.31)
Real space refinement (round 2; 3.25A)
Model RS2 (R conv = 0.30)
Rigid body refinement
Refined model: empty capsids
(R conv = 0.21; Wu et al., 1993)
Real space refined into map of empty capids.
Improve stereochemical restraints
Local real space refinement
Phase calculation; Resolution extended to 2.9A
Phase refinement, map improvement by 10 cycles of symmetry averaging.
Real space refinement (round 3; 2.9A)
Model RS3 (R free = 0.28)
Minor remodeling (R free = 0.28)
Real space refinement (round 4; 2.9A)
Model RS4 (R free = 0.28)

Fig. 1. Steps in the refinement of DNA-containing canine parvovirus. For comparative purposes, the R factors were all calculated using ~1000 reflections selected at random between 5 and 3.25 Å resolutions, for which $F > 5\sigma_F$. Table 10 shows R factors calculated to higher resolution and without additional F/σ_F selection.

Table 3. *Refinement statistics for round 1*

$R^{\rm conv}$ were calculated with subsets of ~500 or ~1000 reflections between ∞ and 3.25 Å resolution and with $F > 5\sigma_F$. $\Delta\alpha$ is the mean absolute phase difference between the phases used to calculate the map and those calculated from the real-space refined atomic model. $R^{\rm ED}$ is defined (Chapman, 1994) as $R^{\rm ED} = \{\sum_{xyz}[S\rho_o(xyz) + k - \rho_c(xyz)]\}/\{\frac{1}{2}\sum_{xyz}[S\rho_o(xyz) + k + \rho_c(xyz)]\}$, where the *xyz* summation includes all grid points within the refining radius ($r_{\rm ref} = 1.6$ Å) of any atom, and ρ_c includes the contributions of all atoms within a sphere of a larger cut-off radius ($r_{\rm calc} = 3.25$ Å) of each grid point. Note that the value of $R^{\rm ED}$ was incorrectly defined by Chapman (1994) and is corrected here.

							R.m.s. geometrical error		
	No. of				R.m.s. shift		Bond	Bond	Torsion
	long			$\Delta\alpha$	Batch	Overall	length	angle	angle
Batch	cycles	$R^{\rm conv}$	$R^{\rm ED}$	(°)	(Å)	(Å)	(Å)	(°)	(°)
0	—	0.341	0.690	26.6			0.021	2.32	19.9
1	16	0.307	0.555	20.5			0.035	3.32	19.8
Middle of 2				0.364				0.313	20.8
27.0									
2	51	0.309	0.555	19.5	0.35	0.55	0.012	2.38	22.3
Middle of 3			0.290	0.452	17.3			0.059	7.94
22.1									
3	51	0.306	0.548	19.0	0.15	0.62	0.012	2.55	23.2
3b	14	0.303	0.539	190			0.018	2.48	21.3
4	13	0.310	0.556	19.5	0.21	0.57	0.017	3.28	11.3

Table 4. *Refinement statistics for round 2*

The number of poor contacts per protomer is given in parentheses, following the r.m.s. contact error. See Table 3 for definitions of $R^{\rm conv}$, $R^{\rm ED}$ and $\Delta\alpha$.

						R.m.s. geometrical error			
	No. of					Bond	Bond	Torsion	Non-bonded
	long			$\Delta\alpha$	R.m.s. shift	length	angle	angle	contact
Batch	cycles	$R^{\rm conv}$	$R^{\rm ED}$	(°)	(Å)	(Å)	(°)	(°)	[Å (number)]
Start		0.312	0.574	19.4			2.62	14.6	0.15 (103)
1	14	0.298	0.522	17.8	0.17	0.022	2.60	15.2	0.12 (110)

Table 5. *Weights used for refinement*

				Geometry				*B*-factor
	RSRef	Bond length	Angle	Torsion	Contact	Trigonal planes	General planes	variation
Range (rounds 2* to 4)	70–103	1–1.7	1.7–2.0	0.5–1.0	1.0–3.8	1.0	3.0	0.4–0.7
Final	103	1.45	2.0	1.0	3.0	1.0	3.0	0.6

* Many weighting schemes were tried in round 1.

Table 6. *Rigid-body refinement*

	Translation from original position (fractional units)			Rotation (°) about the three orthogonalized (*XYZ*) axes				Unit-cell scale	Free *R* factor, $R_T^{\rm free}$ (No. of reflections)	
Cycle	*x*	*z*	Total	φ_x	φ_y	φ_z	Total*	factor†	(500)	(1000)
0									0.3055	
1	0.00030	0.00023		0.050	0.075	0.052			0.2968	
2	0.00053	0.00048		0.045	0.055	0.047				0.3024
3	0.00047	0.00025		0.045	0.052	0.032				0.3005
4	0.00042	0.00023	0.00048	0.045	0.042	0.032	0.069	1.0003		0.2993

* Total is defined here as $(\varphi_X + \varphi_Y + \varphi_Z)^{1/2}$. † The scale factor by which the unit-cell lengths were isotropically magnified to give the lowest $R_T^{\rm free}$.

non-bonded contact; secondly, the residues whose main-chain torsion angles fell outside the allowable regions of a Ramachandran plot (Ramachandran, Ramakrishnan & Sasisekharan, 1963; Morris, MacArthur, Hutchinson & Thornton, 1993); thirdly, the 50 residues with worst fit to the electron density according to a residue-by-residue *R* factor calculated by *RSRef* that is conceptually similar to the *R* factor of Jones *et al.* (1991). The nucleic acid was rebuilt, because, with refinement improving the phosphoribose backbone conformation, better fitting bases could be chosen. Two additional metal ions associated with the protein were also defined.

By this stage, the refinement of the empty capsids had been completed (Wu *et al.*, 1993). Superficially, the structures looked similar with major differences being only on interior and exterior surfaces where interactions

Table 7. *Refinement statistics for round 3*

						R.m.s. stereochemical error			
	No. of long			R.m.s. ΔB	R.m.s. shift	Length	Angle	Torsion	Contacts
Batch	cycles	$R_T^{\rm free}$	$R^{\rm ED}$	(Å^2)	(Å)	(Å)	(°)	(°)	[Å (No.)]
	Start	0.290	0.498	0.0	N/A	0.022	2.60	15.2	0.12 (110)
1	10	0.290	0.477	0.0	0.11	0.022	3.22	15.2	0.10 (112)
2	6		0.446	0.0	0.05	0.019	2.97	15.4	0.11 (145)
3	6	0.282	0.415	2.5	N/A	0.019	2.97	15.4	0.11 (145)
4.2	8	0.280	0.405	3.0	0.09	0.017	2.84	15.7	0.10 (177)
5	Start	0.282	0.429	N/A	N/A	0.018	2.85	15.7	0.32 (197)
5	8	0.282	0.425	N/A	0.06	0.017	2.78	15.7	0.11 (179)
6	8		0.428	2.0	N/A	0.018	2.81	15.7	0.13 (170)
7	5		0.426	2.3	N/A	0.018	2.81	15.7	0.13 (170)
8	3	0.283	0.426	2.3	N/A	0.017	2.76	15.7	0.20 (196)
9	7	0.282	0.426	2.3	N/A	0.017	2.74	15.7	0.12 (208)

Table 8. *Optimization of B-factor variation*

ΔB is the root-mean-square variation in temperature factors between covalently linked atoms. The appropriate value (2.4 Å^2) for minimal $R_T^{\rm free}$ was determined by interpolation and corresponded to a weight of 0.6.

								R.m.s. stereochemical error		
	No. of long			R.m.s. ΔB	Weight	R.m.s. shift	Length	Angle	Torsion	Contacts
Batch	cycles	$R_T^{\rm free}$	$R^{\rm ED}$	(Å^2)	on ΔB	(Å)	(Å)	(°)	(°)	[Å (No.)]
4.1	8	0.282	0.418	1.5	1.0	0.09	0.018	2.83	15.7	0.10 (143)
4.2	8	0.280	0.405	3.0	0.4	0.09	0.017	2.84	15.7	0.10 (177)
4.3	8	0.281	0.393	4.4	0.2	0.09	0.017	2.68	15.7	0.10 (165)
4.4	8	0.283	0.385	6.6	0.1	0.09	0.014	2.55	15.7	0.09 (162)
4.5	8	0.287	0.379	9.3	0.04	0.09	0.014	2.49	15.5	0.09 (157)

Table 9. *Refinement statistics for round 4*

						R.m.s. stereochemical error			
	No. of long			R.m.s. ΔB	R.m.s. shift	Length	Angle	Torsion	Contacts
Batch	cycles	$R_T^{\rm free}$	$R^{\rm ED}$	(Å^2)	(Å)	(Å)	(°)	(°)	[Å (No.)]
	Start	0.282	0.419	3.0	N/A	0.018	2.79	14.8	0.11 (143)
1	7		0.401	3.0	0.06	0.019	2.93	15.3	0.11 (158)
2	3		0.404	2.7	N/A	0.019	2.93	15.3	0.11 (158)
3	5	0.279	0.402	2.5	0.08	0.018	2.74	15.3	0.11 (166)

Table 10. *Final free R factors*

The free R factors shown here were calculated from 7461 randomly chosen reflections without any F/σ_F selection. These reflections had not been used for map calculations or refinements.

Resolution range (Å)	10.0–4.1	4.1–3.9	3.6–3.5	10.0–3.5	3.35–3.25	10.0–3.25	3.0–2.9	10.0–2.9
$R_T^{\rm free}$	0.25	0.29	0.33	0.266	0.36	0.275	0.43	0.291
Fraction of theoretically possible observed reflections with $I > 2\sigma_I$	0.83	0.69	0.56		0.44		0.22	

with DNA or neighboring particles differed. Comparison of the fully refined model of the empty capsid (conventional $R^{\rm conv} = 0.21$) with the map of the full capsid, either by inspection or calculation of real-space residue R factors (Jones *et al.*, 1991), showed that the partially refined full capsid model fitted the electron density better in all but a few places. The empty capsid structure was of use, however, after it was refined with 15 cycles of real-space refinement against the map of the full capsid, during which the real-space R factor ($R^{\rm ED}$; Chapman, 1994) fell from 0.62 to 0.52. In several places, totaling 75 residues, this 'hybrid' model fit the density well, had good stereochemistry, and was used to replace the corresponding parts of the partially refined full capsid model (RS1). The hybrid model continued to be of occasional use in later rounds of remodeling.

Rebuilding was facilitated through local application of real-space refinement run as a macro from *O* (Jones *et al.*, 1991). After crude rebuilding, 'zones' of about five residues at a time were refined. All residues with atoms in a box enclosing the zone were refined, and those within an additional boundary region were used for electron density and stereochemical calculations, but were fixed for refinement. Generally three to five cycles of refinement were required, each taking about 75 s. Weights were varied to drive the model towards good geometry or good fit, as appropriate.

Inspection of model RS1 revealed repeated instances of stereochemical errors, suggesting that some inappropriate restraints had been used during refinement. The list of restraints distributed with *TNT* (Tronrud *et al.*, 1987) was modified as follows and the modified restraints were used for rounds 2 through 4 and for all local refinements. Firstly, bond-angle restraints were added to keep pyrimidine N1 and purine N9 atoms trigonal planar. Secondly, the target value of the phosphoribose backbone torsion angle, ζ, was changed from one typical for B-DNA (260°) to one intermediate between A and B forms (290°; Saenger, 1984). Thirdly, the restraint on the peptide torsion angle, φ, was changed. The default *TNT* targets favor the selection of one of three rotamers, 120° apart. This is inconsistent with the Ramachandran plot (Ramachandran *et al.*, 1963) in which φ has two favored values. The default *TNT* weight is zero for φ and ψ, meaning that in most refinements, no restraints are applied. Starting at such low (3.25 Å) resolution with CPV, it was appropriate to restrain the main-chain torsion angles. The default targets gave a trimodal distribution of torsion φ, with peaks either side of the most favored region ($\varphi \simeq -105°$, $\psi \simeq 135°$; Ramachandran *et al.*, 1963; Morris *et al.*, 1993). Subsequent to round 1, the targets were changed to be consistent with Ramachandran *et al.*, with optimal values for φ defined at 75° and −105°, and ψ defined at 135° and −55°.

Following the rebuilding of RS1, the model was refined with 14 cycles in a second round of real-space refinement (Table 4). Weights were chosen (Table 5) to retain good stereochemistry. For the last couple of cycles, the weight on non-bonded contacts was increased. Additional batches (not shown) failed to further improve the model. The refined model was named RS2.

2.2. *Rigid-body refinement*

Errors in the position or orientation of the particle within the unit cell or in the cell parameters cause smearing and distortion of the electron density upon icosahedral averaging. Models fitted to such maps will have poorer *R* factors. The position and orientation of the particle were optimized in four cycles of rigid-body refinement. The *y* coordinate of the reference particle can be arbitrarily fixed in space group $P2_1$. The largest component of error in unit-cell dimensions, following post-refinement (Rossmann, 1985), is likely to be an isotropic magnification resulting from error in the assumed wavelength of synchrotron radiation. A scale constant for isotropic magnification of the unit cell lengths was introduced on the fourth cycle. In each cycle, the coordinates for a complete virus particle were translated and rotated in each direction before calculating structure amplitudes to search for the lowest *R* factor. To reduce computation, calculated structure amplitudes were compared to 500 (cycles 1 and 2) or 1000 (cycles 3 and 4) randomly selected observed amplitudes out of the more significant data ($F > 3\sigma_F$) between 5 and 3.25 Å resolution. The process converged in four cycles (Table 6).

2.3. *Further phase refinement and extension*

Earlier attempts to extend the phases beyond 3.25 Å resolution by electron-density averaging had been unsuc ssful (see section on initial structure determination). If a slightly incorrect particle position or orientation had contributed to this failure, this could now be remedied. It is also possible that some phases had been unable to move from incorrect values, such as Babinet opposites of the correct phases, because they were self-consistent within a local region of reciprocal space (Välegard, Liljas, Fridborg & Unge, 1991; McKenna *et al.*, 1992). However, by starting with phases calculated from an atomic model, such inconsistencies can be avoided.

The observed data set was randomly split into two unequal groups. 576 286 reflections (98.7%) were used for phase refinement and map calculation, while 7461 reflections (1.3%) were saved for future calculation of the free *R* factor ($R_T^{\rm free}$; Brünger, 1992) and were omitted from further map calculations. A complete set of ~1.2 million structure factors was calculated to 2.8 Å resolution from the ~300 000 atoms per asymmetric unit, a task that can be accomplished in about 4 h by FFT methods (Ten Eyck, 1973, 1977) using an Evans and Sutherland graphics workstation. Phases calculated from the model were paired with the observed magnitudes. For reflections that had not been observed, structure factors were initially calculated from the model, but these were replaced in each cycle of electron-density averaging phase refinement by those calculated by back transformation of the averaged map of the previous cycle. At high resolution, the accuracy and the fraction of observed reflections decrease, reducing the power of phase refinement (Arnold & Rossmann, 1988). Hence, phase extension was limited to 2.9 Å resolution, beyond which < 20% of the reflections had been observed.

The electron-density averaging program described by Rossmann *et al.* (1992) was used with figure of merit weighted maps computed on a 1.1 Å spaced grid. The particle envelope was defined by concentric spheres with radii of 69.4 and 146.2 Å and planes midway between adjacent particles. The radii were chosen with a 3.5 Å margin to include every atomic position. Ten cycles were required to reach convergence.

2.4. *Further real-space model refinement*

In round 3, model 2 was refined against a new map calculated from the refined phases to 2.9 Å resolution.

The fit of the starting model to the new map was slightly better ($R^{\text{ED}} = 0.50$) than the previous map ($R^{\text{ED}} = 0.52$). Batch 1 (Table 7) was a conservative refinement in which small atomic movements (r.m.s. of 0.1 Å) improved the fit to the new map, dropping R^{ED} from 0.50 to 0.48, but without improvement in R_T^{free}. Batch 2 started at a slightly lower $R^{\text{ED}} = 0.454$, because, consistent with the higher resolution of the new map, the cut-off for calculation of overlapping atomic electron densities (r_{calc}) was decreased from 3.25 to 3.0 Å, increasing the speed of refinement.

Independent isotropic temperature factors were first introduced in batch 3 (Table 7). Additional restraints were added to the defaults of *TNT*, so that, as in *PROLSQ*, the differences in B factors between side-chain atoms and covalently linked main-chain atoms were restrained. Brünger (1992) showed that models can be 'overfit' to structure amplitudes by adjusting the atomic parameters with restraints that are too weak. To avoid the danger of similarly overfitting the model to the map, the appropriate variation for CPV ($\Delta B = 2.4\,\text{Å}^2$) was found by searching for the lowest R_T^{free} among parallel refinements that used different weights on the B-factor restraint, but were otherwise identical (Table 8). The B-factor weight was adjusted on subsequent refinement to keep the r.m.s. ΔB close to the target of 2.4 Å^2.

In batch 5, 75 water molecules were added to the model, mostly on the outer and inner surfaces of the capsid. These had been positioned during the rebuilding of RS1 in unmodeled electron density where there was reasonable hydrogen bonding geometry, but had hitherto not been used in refinement. In batches 5 through 9, first positions, then temperature factors were refined. Eight cycles of positional refinement, with fixed temperature factors, were performed in batch 5. Batches 6–9 refined both positions and temperature factors, although protein and nucleic acid changed little. Ultimately, the water molecules were discarded, because there were several indications that they were not justified at 2.9 Å resolution. Firstly, during refinement, the fit to the map (R^{ED}) improved slightly, but R_T^{free} did not. Secondly, many water molecules had moved by more than 0.7 Å. Thirdly, there was little correspondence in the positions of waters in the DNA-containing and empty (Wu & Rossmann, 1993) models.

In batch 7, B factors for two of the metal ions that were modeled as Mg^{2+} reached their lower limits (0), suggesting that they might be heavier ions. Further refinement as Ca^{2+} led to B factors of 50 and 70 Å^2, slightly higher than the surrounding DNA ($\langle B \rangle \simeq 42\,\text{Å}^2$), suggesting that these sites might be occupied by a mixture of heavy and light ions. As two of the ions are an integral part of the DNA structure, it is likely that one of the two ions is present for every DNA fragment that is seen.

The product of this refinement, model RS3, was modified interactively using *O* (Jones *et al.*, 1991), and using *ProCheck* (Laskowski, MacArthur, Moss & Thornton, 1993) to highlight poor parts of the model. The subsequently modified model was then real-space refined in round 4 (Table 9). Batches 1 and 2 refined the B factors and positions alternately, before combined refinement in batch 3. The product of the fourth round of refinement was named RS4.

3. Results

3.1. *Rigid-body refinement*

The position and orientation converge quickly (Table 6). The overall shift in position was about 0.13 Å and orientation was 0.07°. Assuming that the 'average' atom is 108 Å from the viral center (Chapman *et al.*, 1992), the average atomic movement was 0.18 Å. Isotropic scaling of the unit-cell lengths gave a minimal R factor with a 0.03% increase in cell dimensions, corresponding to a change of roughly 0.1 Å to $a = 263.21$, $b = 349.01$, $c = 267.32$ Å. This would correspond to an error of 0.00003 Å in wavelength calibration of the synchrotron X-radiation.

In the initial structure determination of CPV (Tsao, Chapman, Wu *et al.*, 1992), the electron-density map was found to be of the left-handed enantiomer. Therefore, Tsao *et al.* built the protein model into a map transformed by $q' = -q$ to make it right-handed within the orthogonal system (P,Q,R) defined with respect to the usual symmetry axis. The particle position o_L, deorthogonalization matrix $[\alpha_L]$, icosahedral point group and orientation matrix $[\rho_L]$ (that aligns three orthogonal icosahedral diad axes (P,Q,R) with the orthogonalized cell axes] that were determined by Tsao *et al.* are all consistent with the untransformed, left-handed enantiomer. The local $q' = -q$ transformation does not apply throughout the unit cell, complicating structure-factor calculation. Thus, for refinement, a right-handed enantiomer was chosen. By calculating fractional coordinates, a, in the following way, the same non-crystallographic asymmetric unit and 60 icosahedral symmetry rotation matrices, $[\mathbf{R}_n]$, may be used as previously deposited with the Brookhaven Protein Data Bank, and structure factors calculated from these coordinates are compatible with a data set with the negatives of the experimental phases of Tsao *et al.* (1992),

$$\begin{aligned} a &= [\alpha_R][\rho_R][\mathbf{R}_n]p + o_R \\ &= \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} [\alpha_L][\rho_L] \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} [\mathbf{R}_n]p \\ &\quad + \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} o_L. \end{aligned} \qquad (3)$$

$[\alpha_R]$, a deorthogonalization matrix, operates on right-handed orthogonal coordinates, but differs from convention (Rossmann & Blow, 1962) by aligning orthogonalized axes (X, Y, Z) in the opposite directions of the corresponding crystal axes, (A, B, C) (*i.e.* $Y \parallel -B$, $X \parallel -A \wedge B$ instead of $Y \parallel B, X \parallel A \wedge B$). This unconventional alignment could have been avoided by re-indexing the reflections to change. the directions of the crystal axes, but this would have introduced further confusion. The values of the matrices and o_R, the particle center position, were updated from the results of the fourth cycle of rigid-body refinement,

$$o_R = (-0.25392, 0.0, -0.24723);$$

$$[\rho_R] = \begin{pmatrix} 0.5793806 & -0.0105097 & -0.8149953 \\ 0.0300839 & -0.9989412 & 0.0342912 \\ 0.8144993 & 0.0444241 & 0.5785251 \end{pmatrix};$$

$$[\alpha_R] = \begin{pmatrix} -0.0037996 & 0.0000000 & 0.0000000 \\ 0.0000000 & -0.0028652 & 0.0000000 \\ -0.0000534 & 0.0000000 & -0.0037408 \end{pmatrix}. \quad (4)$$

3.2. *Phase refinement*

The model phases used to start phase refinement by symmetry averaging differed from the previously reported phases (Tsao, Chapman, Wu *et al.*, 1992) by an average of 25° for reflections between ∞ and 3.25 Å resolution (the resolution used for the initial structure determination). The mean correlation coefficient between observed structure magnitudes and those calculated by back-transformation of the initial symmetry-averaged map was 0.6603. It rose quickly and smoothly (Fig. 2) to values very similar to those of the earlier refinement. The average phase change was 22°, but the mean phase difference with respect to the phases of Tsao *et al.* had diminished to only 12°. Phases at 3.25 Å resolution changed by 41° and the correlation coefficient was improved (Fig. 2), indicating a small improvement in the quality of the phases at high resolution. Overall, however, there was little improvement. Subjective comparisons of the previous map and one calculated with the re-refined phases confirmed that there was some improvement, as judged, for example, by the appearance of additional bulges for carbonyl O atoms. However, the improvement was small and suggested that additional rigid-body and phase refinement would not be worthwhile.

3.3. *Fit of models to electron density and refinement statistics*

The improvement in the fit of refined models to the electron density is shown for a typical region in Fig. 3. The biggest improvements were made in the first round, between the starting model and RS1. For example, changes in the backbone conformation permitted the side chains of Trp214 and Arg216 (Fig. 3) to be accommodated completely within the electron density. The changes on the next round, from RS1 to RS2, were more modest. The backbone adjusted itself to move the carbonyl O atoms a bit closer to bulges in the electron density. Comparison of the two electron-density maps shows small improvements. Small errors in the electron density, such as those that can be seen in the side-chain

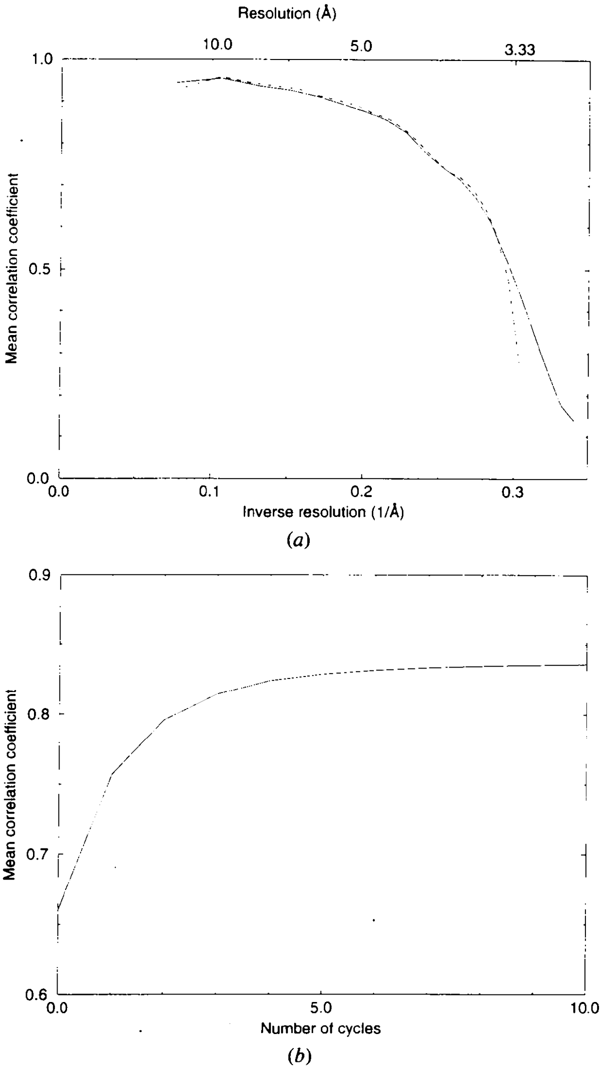

Fig. 2. Phase refinement. (*a*) The mean correlation coefficient between observed and back-transformed structure-factor magnitudes from infinite to 3.25 Å resolution shows rapid and smooth convergence, starting with phases calculated from the atomic model. (*b*) Final correlation coefficients plotted as a function of resolution. The statistics resulting from the phase refinement of the initial structure determination (3.25 Å; Tsao, Chapman, Wu *et al.*, 1992, dotted line) are compared to the results of the phase refinement by electron-density averaging which was started with phases to 2.9 Å resolution calculated from atomic model RS2.

densities of Thr217 and Trp214 (Fig. 3), were removed in the new map. In some areas, the definition of the carbonyl O atoms was improved slightly.

The final free R factor for all data to 2.9 Å resolution was 0.291. R factors for selected resolution ranges are shown in Table 10. Stereochemical statistics for model RS4 were calculated by *TNT*'s *Geometry* routine (Tronrud *et al.*, 1987) and the program *ProCheck* (Laskowski *et al.*, 1993). They are listed in Table 11 where they are compared to corresponding statistics from other well refined protein structures, showing that the stereochemistry is as good as usually found in other ~2.9 Å resolution structures. The main-chain torsion angles show a good distribution (Fig. 4).

Table 11. *Stereochemical statistics for final model, RS4*

All lines except (*e*) and (*i*) show the root-mean-square deviations (r.m.s.d.) from ideal stereochemical values. The statistics for lines (*a*)–(*d*), (*f*) and (*g*) were calculated using *TNT*'s *Geometry* (Tronrud *et al.*, 1987) with typical values for comparison taken from four structures refined at about 2 Å and listed by Tronrud *et al.* (1987). The statistics for lines (*e*) and (*j*)–(*m*) were calculated and compared to database structures refined at about 2.9 Å using the program *ProCheck* (Laskowski *et al.*, 1993).

	Parameter	CPV model RS4	Typical value
(*a*)	R.m.s.d. {bond length} (Å)	0.018	0.02
(*b*)	R.m.s.d. {bond angle} (°)	2.7	2.9
(*c*)	R.m.s.d. {torsion angle} (°)	15.4	
(*d*)	R.m.s.d. {poor contact} (Å)	0.010	
(*e*)	No. of bad contacts/100 residues	3.9	17.9 ± 10.0
(*f*)	R.m.s.d. {trigonal planes} (Å)	0.016	
(*g*)	R.m.s.d. {general planes} (Å)	0.024	
(*h*)	R.m.s.d. {ζ} (Cα chirality) (°)	3.2	3.1 ± 1.6
(*i*)	Residues with most favored (φ, ψ) angles (%)	80	69
(*j*)	R.m.s.d. {ω} (peptide) (°)	3.6	6.0 ± 3.0
(*k*)	R.m.s.d. {hydrogen-bond energy} (kcal mol^{-1})	1.0	1.0 ± 0.2
(*l*)	R.m.s.d. {χ_1} (side-chain torsion angle) (°)	13.6	25.1 ± 4.8
(*m*)	R.m.s.d. {χ_2} (side-chain torsion angle) (°)	15.2	25.3 ± 5.0

4. Discussion

4.1. *R factors and the quality of the final structure*

Refinement by the new program, *RSRef*, was first tested by re-refining the structure of the empty CPV capsid. Real-space refinement was able to produce a model that was at least as good as that produced through reciprocal-space refinement (Chapman, 1994, *cf.* Wu & Rossmann, 1993) as judged through conventional reciprocal space R factors. Real-space refinement of both the empty capsid and of the DNA-containing capsid produced models with similarly good fits to their respective electron densities (as judged by R^{ED}) that

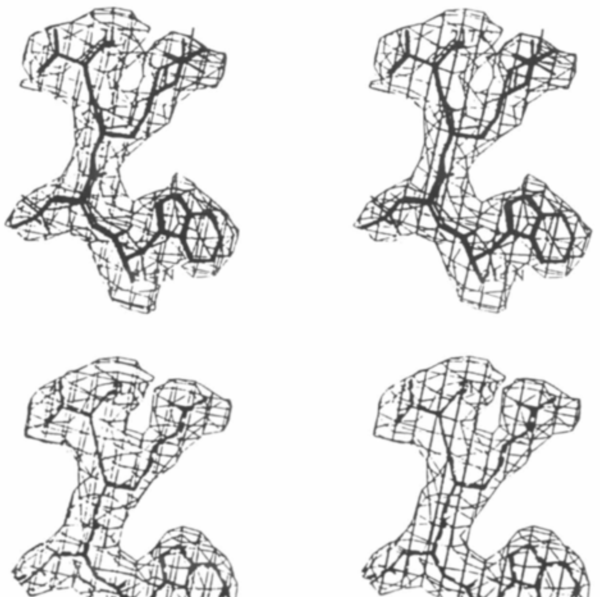

Fig. 3. Fit of the model to electron density. The upper stereogram shows residues 214–217 of the starting model (thin lines), model RS1 (medium) and model RS2 (thick lines) superimposed upon the electron-density map of Tsao *et al.* (1991). The starting model was refined against this map to yield model RS1, which after some manual rebuilding was further refined to RS2. The lower stereogram illustrates the later stages of refinement, showing RS2 (dashed lines) and RS4 (solid lines) superimposed on the final map. The final map was based on phases calculated from RS2 and then refined by 60-fold symmetry averaging. This figure was prepared using the program *O* (Jones *et al.*, 1991).

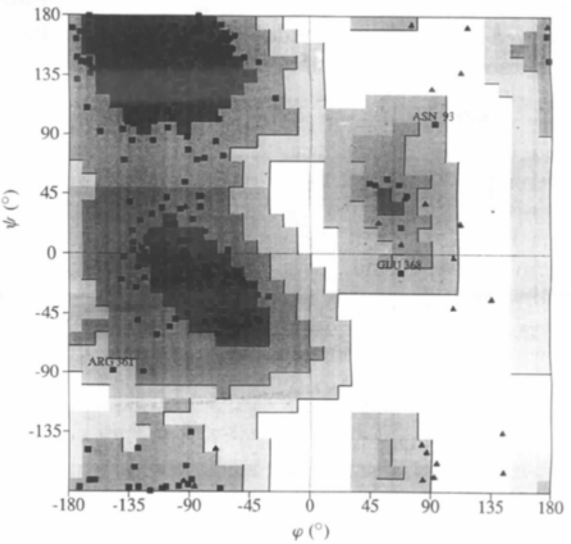

Fig. 4. Ramachandran plot (Ramachandran *et al.*, 1963) showing the distribution of main-chain torsion angles for model RS4. Glycines are shown as triangles, and other amino acids as squares. This plot was made using the program *ProCheck* (Laskowski *et al.*, 1993). The most favored 'core' regions, defined by Morris *et al.* (1993), are shaded darkly. The three non-glycine residues that fall outside the 'allowed' region (medium shading), are labeled and all fall within the 'generously allowed' regions.

were better than the fit of the reciprocal-space refined empty capsid structure to its electron density. Thus, at first sight, it was disappointing that the R factor for the real-space refined DNA-containing capsid was 6% higher than that of the reciprocal-space refined empty capsid, because this suggested that real-space refinement might not always be as good as reciprocal-space refinement. However, as demonstrated in the *Appendix*, the difference in R factors can be fully accounted for by the presence of DNA in the full capsid (87% of which is disordered), and by differences between free and conventional R factors. Cross-validation R factors (see *Appendix*) show that the protein of the DNA-containing capsid is modeled at least as well as that of the empty capsid. The disordered DNA that cannot be modeled elevates the R factor significantly. Even so, the free R factor, $R_T^{\rm free} = 0.29$ for the DNA-containing capsid compares favorably with those of protein structures, refined at ~3 Å, for which usually $R_T^{\rm free} > 0.3$ (Brünger, 1992).

4.2. *Refinement of position, orientation and unit-cell constants*

The change in particle orientation determined by the reciprocal-space method described here was 1.5 times larger than the error estimated by Tsao, Chapman, Wu *et al.* (1992). The error had been estimated from the r.m.s. deviation between rotation-function peaks and the directions expected of a standard icosahedron. Nevertheless, the error in the original orientation, corresponding to a movement of 0.2 Å for an average atom, is only 1/20th of the 4.3 Å resolution limit used in the rotation function (Tsao, Chapman & Rossmann, 1992).

In the initial structure determination, the final position of the particle was determined as the one that gave the lowest r.m.s. deviation of electron densities from symmetry-equivalent positions. As this is a real-space determination, there is the potential of bias towards the current position through the phases. Cycles 2 through 6 of the position refinement (Tsao, Chapman, Wu *et al.*, 1992) changed the position in roughly the same direction, with a final step of 0.08 Å. The reciprocal-space rigid-body refinement, described here, moved the particle position in a similar direction, by about 0.13 Å. This confirms that there was some negative feedback bias in the method of Tsao, Chapman, Wu *et al.* (1992), but that the bias was modest.

4.3. *Phase refinement*

There was negligible improvement of the phases or map after re-initiating phase refinement with phases calculated from the intermediate model, RS2. This suggests that the resolution limit of phase extension was not imposed by methodology [different averaging programs were used by Tsao, Chapman, Wu *et al.* (1992) and for the work described here]; by errors in particle position or orientation; or by phase errors accumulated through extension. Rather it indicates that the limit is due to the weakness of the data at high resolution, that is manifest both in terms of structure-amplitude error and low fraction of data that has been collected (Table 10).

4.4. *Use of second derivatives*

Tronrud (1992) found that the convergence properties of reciprocal-space refinement are improved with use of the second derivatives to optimize calculation of the shift vector. There were modest improvements in the convergence properties of the real-space method during the later stages of a round of refinement. However, at the start of a round, if either the stereochemistry or fit to the electron density were poor, faster convergence was achieved without second derivatives. The reason is likely to be that the functional forms of the stereochemical and electron-density terms are so different – parabolic for the stereochemical terms and the Fourier transform of a truncated Gaussian for electron-density terms. The derivatives of the two types of restraints may be of similar shape when close to convergence, but elsewhere the stereochemical terms become larger and larger, while those of the electron density tail off towards zero. Although it is possible to find a scale constant (weight) that approximately scales the first derivatives of the stereochemical and electron-density terms, it is not possible to find a single scale constant to scale both first and second derivatives for both types of terms. In practice, when second derivatives are used far from convergence, there is a strong tendency for atomic shifts to be unreasonably small (asymptotic approach) or too large (oscillatory approach), depending on the exact choice of weights.

4.5. *General applicability and phase limitations*

The structure determination of viral capsids differs from that of most proteins in that the phases derived from experimental data have greater accuracy (Arnold & Rossmann, 1988) and are independent of the assumption of any kind of atomic parameters. In general, they are therefore likely to be more accurate than phases calculated from atomic models. Use of calculated phases can bias the results towards the initial model, even when omit maps are used (Hodel, Kim & Brünger, 1992) unless, as with virus capsids, there are strong independent constraints to be imposed subsequently in phase refinement. Thus, real-space methods are advantageous for problems with high non-crystallographic redundancy, but are unlikely to replace reciprocal-space methods for the final refinements of structures lacking extensive non-crystallographic symmetry, including most proteins.

Real-space refinement can also be used to advantage at intermediate steps during protein (or virus) structure determination, to improve the fit of models to electron density. For instance, local refinement greatly facilitated the rebuilding of intermediate CPV models. Real-space refinement can be run, while model building, on local regions of a structure within a few minutes, substantially reducing the time required to build a model that is a good starting point for further refinement by reciprocal-space methods. In this mode, *RSRef* (Chapman, 1994) is conceptually similar to *RSR* (Jones, 1978; Jones *et al.*, 1991), but *RSRef* accounts for the resolution of the map and restrains atoms to idealized geometries (rather than searching with rigid fragments), including non-bonded contacts and single-bond torsion angles.

6. Concluding remarks

Canine parvovirus has been refined successfully by a stereochemically restrained, resolution-dependent real-space refinement method. Refinement is at least 50 times faster than previous methods of virus structure refinement, while using an Evans and Sutherland workstation with ~3 MBytes of memory available for program execution. The method can also be used to easily refine the structures of mutants and complexes of viruses with ligands (*e.g.* antiviral drugs) by confining interest to the region near the altered site.

This work was supported by an NSF grant to Michael G. Rossmann and by grants from the Lucille P. Markey Foundation for the development of structural biology at Purdue University and Florida State University. We are grateful to our colleagues, Jun Tsao, Mavis Agbandje, Walter Keller and Hao Wu, at earlier stages of the CPV structure determination and to Hao Wu for access to the CPV tetragonal structure prior to publication.

7. APPENDIX
Written by Michael Chapman

Cross-validation R factors and their use in comparing the qualities of refined models for the DNA-containing and empty capsids of canine parvovirus

This paper describes the refinement of the canine parvovirus (CPV) DNA-containing capsid by a new real-space method (Chapman, 1994). Evaluation of the quality of the refined model is important in assessing the new method, and is possible because the structure of the empty capsid had previously been refined by conventional reciprocal-space methods (Wu *et al.*, 1993). Direct comparison of R factors is not possible, because the structures differ in the presence/absence of (mostly unmodeled) DNA. Furthermore, the empty capsid model was evaluated using a conventional R factor which has recently been shown to underestimate the error of a model (Brünger, 1992) . Calculation of a free R factor for a previously refined virus structure is impractical, because this involves computationally intense re-refinement, omitting the subset of the data that will be used for evaluation (Brünger, 1992). Similar situations are likely to arise in the future, where there is a need to compare the quality of a recently determined variant structure to that of a previously refined parent.

Fortunately, the existence of data sets for related structures provides an opportunity for cross-validation analogous to that suggested by Brünger (1992), providing that the structures have been refined independently. The experimental data of the variant have not been used in the refinement of the parent, and *vice versa*, and can be used for an independent evaluation similar to a free R factor. If the models and data sets are of comparable quality, their agreement with the data of the other structure should be similar. Both cross-validation R factors will be higher than free R factors, because of expected conformational differences, but will be an independent evaluation of relative quality. If the conformational differences are characterized and affect a small proportion of the molecule, hybrid models can be constructed where the appropriate regions of the parent model are substituted by the corresponding regions of the variant before comparison to the experimental data of the variant (and *vice versa*). The resulting cross-validation R factors will have similar magnitude to free R factors and provide an independent evaluation of the quality of the parts of the model that were not expected to change. This is exactly what is required to assess the qualities of refinements of the DNA-containing and empty CPV capsids.

Let ${}_aR_{db}$ be an R factor calculated by comparing model a to the experimental data for molecule b. Let e denote the empty capsid; f, the full capsid containing DNA; e', a modified model of the empty capsid to which the DNA model has been added, in which 22 amino acids adjacent to the DNA have been substituted by those of the full capsid and which has been appropriately oriented and positioned in the unit cell of the full capsid; and f', a model of the full capsid from which the DNA has been removed, with the same 22 amino acids changed to those of the empty structure and placed in the unit cell of the empty capsid. Water molecules had not been modeled for the full capsid and were removed from the empty capsid model. R factors were calculated with ~1000 randomly selected reflections below 3 Å resolution,

$$ {}_{f'}R_{\mathrm{d}e} = 0.28; {}_{e'}R_{\mathrm{d}f} = 0.31. \tag{5}$$

This suggests that the model of the full capsid is better. This type of analysis is likely to be valid for most comparisons of structures. However, for CPV a more sophisticated analysis is required, because all R factors calculated against the data for the full capsid are expected to be higher due to the unmodeled DNA.

A list of the four self- and cross-validation R factors that can be calculated for the two CPV structures is given in Table 12, along with the presumed dominant components of the discrepancy between observed and calculated structure magnitudes in each case. These components include: the remaining errors in the full and empty capsid models, (ε_f, ε_e, respectively), errors in the two data sets ($\varepsilon_{{\rm d}f}$, $\varepsilon_{{\rm d}e}$), the contribution of the unmodeled DNA to scattering ($\varepsilon_{\rm DNA}$) and other real differences in the protein conformation relevant to cross comparisons, due, for example, to different crystal packing (ε_Δ). Some of these components such as $\varepsilon_{{\rm d}f}$ and $\varepsilon_{{\rm d}e}$ may be estimated from known parameters ($R_{\rm merge}$). The others, including the errors of the two models, may be approximated by solving the following simultaneous equations that assume that R factors approximate standard errors (Hamilton, 1964),

$$({}_fR^{\rm free}_{T,{\rm d}f})^2 \simeq \varepsilon_f^2 + \varepsilon_{{\rm d}f}^2 + \varepsilon_{\rm DNA}^2 \simeq 0.279^2, \qquad (6)$$

$$({}_eR^{\rm conv}_{A,{\rm d}e} + \Delta^{\rm conv}_{\rm free})^2 \simeq \varepsilon_e^2 + \varepsilon_{{\rm d}e}^2 \simeq (0.231 + \Delta^{\rm conv}_{\rm free})^2, \qquad (7)$$

$$({}_{f'}R^{\rm cross}_{T,{\rm d}e})^2 \simeq \varepsilon_f^2 + \varepsilon_{{\rm d}e}^2 + \varepsilon_\Delta^2 \simeq 0.280^2, \qquad (8)$$

$$({}_{e'}R^{\rm cross}_{T,{\rm d}f})^2 \simeq \varepsilon_e^2 + \varepsilon_{{\rm d}f}^2 + \varepsilon_\Delta^2 + \varepsilon_{\rm DNA} \simeq 0.311^2. \qquad (9)$$

It is also necessary to make one of several possible assumptions about the expected difference, $\Delta^{\rm conv}_{\rm free}$, between R factors calculated with the same data as used in refinement, and those calculated with independent data. If it is generously assumed that there is no systematic difference ($\Delta^{\rm conv}_{\rm free} = 0$), then,

$$\varepsilon_e = 0.201;\ \varepsilon_f = 0.209;\ \varepsilon_\Delta = 0.148;\ \varepsilon_{\rm DNA} = 0.133, \qquad (10)$$

suggesting that the empty capsid model is marginally better than that of the DNA-containing capsid, but that most of the difference in R factors is attributable to the unmodeled DNA.

Table 12. *R factors and sources of error for full and empty CPV capsids*

${}_aR_{{\rm d}b}$ is an R factor comparing the structure amplitudes from model a with the experimental data of molecule b. The R factors shown are either free ($R^{\rm free}_T$; Brünger, 1992), calculated using a test subset, T, of the experimental data that were not used for refinement; conventional ($R^{\rm conv}_A$), calculated using the same set (A) of reflections used for refinement; or cross ($R^{\rm cross}_T$) which are calculated using a set, T, of reflections from a different, but related structure determination, and are, therefore, like the free R factor, independent of the refinement.

R factor	Explanation	Value	Major sources of discrepancy between $\lvert F_o\rvert$ and $\lvert F_c\rvert$
${}_fR^{\rm free}_{T,{\rm d}f}$	Full model *cf.* data for full capsid.	0.279	Error in model of full capsid (ε_f); error in experimental data of full capsid ($\varepsilon_{{\rm d}f}$); contribution to scattering of unmodeled DNA ($\varepsilon_{\rm DNA}$)
${}_eR^{\rm conv}_{A,{\rm d}e}$	Empty model *cf.* data for empty capsid.	0.231	Error in model of empty capsid (ε_e); error in experimental data of empty capsid ($\varepsilon_{{\rm d}e}$)
${}_{f'}R^{\rm cross}_{T,{\rm d}e}$	Modified full model *cf.* data for empty capsid.	0.280	Error in model of full capsid (ε_f); error in experimental data of full capsid ($\varepsilon_{{\rm d}e}$); significant differences between the two structures that were not changed in the modified model, such as due to crystal packing (ε_Δ)
${}_{e'}R^{\rm cross}_{T,{\rm d}f}$	Modified empty model *cf.* data for full capsid.	0.311	Error in model of empty capsid (ε_e); experimental data of full capsid ($\varepsilon_{{\rm d}f}$); error in contribution to scattering of unmodeled DNA ($\varepsilon_{\rm DNA}$); significant differences between the two structures that were not changed in the modified model, such as due to crystal packing (ε_Δ)

How much lower is a conventional R factor expected to be for a virus capsid structure? Brünger (1992) suggested that the lowering of conventional R factors was due to over-fitting of the model to the experimental data. In maximizing the agreement between calculated and observed structure factors, during refinement, a model can be adjusted detrimentally to account for part of the residual difference that is really due to experimental error in the observed data. If the same data are used for model evaluation, an R factor can be lowered by this process without model improvement. Brünger (1992) showed that the difference between conventional and free R factors ($\Delta^{\rm conv}_{\rm free}$) was often about 0.15 for proteins. The magnitude of $\Delta^{\rm conv}_{\rm free}$ should decrease with larger data:parameter ratios and accuracy of observed data. A conventional R factor should, therefore, better approximate a free R factor when the number of data is large (high resolution and/or large unit cell) or when the effective number of independent atomic parameters is small (tight stereochemical restraints, non-crystallographic symmetry, *etc.*). Thus for viruses, with large data sets and high non-crystallographic symmetry, the difference between conventional and free R factors is expected to be smaller than for proteins.

For the CPV empty capsid final model (including waters), $R^{\rm conv}_A = 0.211$ when calculated with reflections used in the last cycle of refinement (Wu *et al.*, 1993), but $R^{\rm conv}_{A'} = 0.225$ when calculated with reflections used in the previous three cycles of refinement. Reflections used in previous cycles are not independent of refinement, so this test gives a lower limit to the difference between $R^{\rm conv}_A$ and $R^{\rm free}_T$ of 0.015. For computational expediency, many virus structures, including the bacteriophage φX174 (McKenna *et al.*, 1992) and Flock House virus (FHV; Cheng *et al.*, 1994), have been refined in reciprocal space against alternating subsets of the observed data. Thus, for the first few cycles there are data which have not been used for refinement, and which may be used for free R-factor calculation. The difference between free and conventional R factors is dependent on the number of reflections used. Refined against 180 000 A and B components of the structure-factor vectors (*cf.* Arnold & Rossmann, 1988), $\Delta^{\rm conv}_{\rm free} \simeq 0$ for the first cycle of FHV refinement, but is significant ($\Delta^{\rm conv}_{\rm free} = 0.045$) when only 14 000 A and B components are used. It also increases with the number of cycles, as demonstrated with the refinement of φX174 against 60000 A and B components. For cycle 1, $\Delta_{\rm free}^{\rm conv} = 0.008$ rising to $\Delta_{\rm free}^{\rm conv} = 0.016$ on the second cycle, as the conventional R factor falls from 0.291 to 0.261. It is not possible to calculate free R factors for subsequent cycles without re-refining, as, by this stage, all of the reflections have been used in refinement. Extrapolation to the final conventional R factor for φX174 suggests that the final $\Delta_{\rm free}^{\rm conv}$ is likely to be about 0.03. For the CPV empty capsid, $\Delta_{\rm free}^{\rm conv}$ may be larger, because of the relatively poor quality and uneven distribution of data that resulted from the large unit-cell dimensions, and because fewer reflections (45000) were used for each cycle (Wu *et al.*, 1993). If it is assumed that $\Delta_{\rm free}^{\rm conv} = 0.03$ then,

$$\varepsilon_e = 0.235; \varepsilon_f = 0.226; \varepsilon_\Delta = 0.120; \varepsilon_{\rm DNA} = 0.103, \qquad (11)$$

suggesting that the model of the DNA-containing capsid is slightly better than that of the empty capsid.

Considering the approximations made in this calculation, there appears to be no significant difference in the qualities of the two models. The difference in the R factors is attributable in roughly equal measure to the difference between conventional and free R factors and to the presence of unmodeled DNA in the full capsid. A similar approach can be used to assess the relative qualities of other pairs of structures when free R-factor calculation is impractical. Usually, simple cross-validation R factors [as in (5)] suffice, but analysis of the separate error components (6)–(9) may be required if a significant part of one structure is unmodeled.

MSC is grateful to R. McKenna and to V. S. Reddy for the calculation of R factors for φX174 and FHV, respectively.

References

Anderson, L. J. & Török, T. J. (1989). *New Engl. J. Med.* **321**, 536–538.

Arnold, E. & Rossmann, M. G. (1988). *Acta Cryst.* A**44**, 270–282.

Berns, K. I. (1990). *Virology*, edited by B. N. Fields, D. M. Knipe, R. M. Chanock, M. S. Hirsch, J. L. Melnick, T. P. Monath & B. Roizman, pp. 1743-1763. New York: Raven Press.

Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 458–460.

Chapman, M. S. (1994). *Acta Cryst.* A**51**, 69–80.

Chapman, M. S., Tsao, J. & Rossmann, M. G. (1992). *Acta Cryst.* A**48**, 301-312.

Chen, Z., Stauffacher, C., Li, Y., Schmidt, T., Bomu, W., Kamer, G., Shanks, M., Lomonossoff, G. & Johnson, J. E. (1989). *Science*, **245**, 154–159.

Cheng, R. H., Reddy, V. S., Olson, N. H., Fisher, A. J., Baker, T. S. & Johnson, J. E. (1994). *Structure*, **2**, 271–282.

Diamond, R. (1971). *Acta Cryst.* A**27**, 436–452.

Fisher, A. J. & Johnson, J. E. (1993). *Nature (London)*, **361**, 176–179.

Hamilton, W. C. (1964). *Statistics in Physical Science*. New York: Ronald Press.

Hendrickson, W. A. (1985). *Methods Enzymol.* **115**, 252–270.

Hodel, A., Kim, S.-H. & Brünger, A. T. (1992). *Acta Cryst.* A**48**, 851–858.

Jones, T. A. (1978). *J. Appl. Cryst.* **11**, 268–272.

Jones, T. A. & Liljas, L. (1984). *Acta Cryst.* A**40**, 50–51.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.

Kurtzman, G., Frickhofen, N., Kimball, J., Jenkins, D. W., Nienhuis, A. W. & Young, N. S. (1989). *New Engl. J. Med.* **321**, 519–523.

Larson, S. B., Koszelak, S., Day, J., Greenwood, J., Dodds, J. A. & McPherson, A. (1993). *Nature (London)*, **361**, 179–182.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.

McKenna, R., Xia, D., Willingmann, P., Ilag, L. L. & Rossmann, M. G. (1992). *Acta Cryst.* B**48**, 499–511.

Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1993). *Proteins*, **12**, 345–364.

Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). *J. Mol. Biol.* **7**, 95–99.

Rossmann, M. G. (1985). *Methods Enzymol.* **114**, 237–280.

Rossmann, M. G. (1990). *Acta Cryst.* A**46**, 73–82.

Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24–31.

Rossmann, M. G., McKenna, R., Tong, L., Xia, D., Dai, J., Wu, H., Choi, H. K. & Lynch, R. E. (1992). *J. Appl. Cryst.* **25**, 166–180.

Saenger, W. (1984). *Principles of Nucleic Acid Structure*. New York: Springer-Verlag.

Silva, A. M. & Rossmann, M. G. (1985). *J. Mol. Biol.* **197**, 69–87.

Stubbs, G. (1989). *Protein–Nucleic Acid Interactions*, edited by W. Saenger & U. Heinemann, pp. 87–109. Boca Raton: CRC Press.

Studdard, M. J. (1990). *Handbook of Parvoviruses*, edited by P. Tijssen, pp. 27–32. Boca Raton: CRC Press.

Ten Eyck, L. F. (1973). *Acta Cryst.* A**29**, 183–191.

Ten Eyck, L. F. (1977). *Acta Cryst.* A**33**, 486–492.

Tronrud, D. E. (1992). *Acta Cryst.* A**48**, 912–916.

Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). *Acta Cryst.* A**43**, 489–501.

Tsao, J., Chapman, M. S., Agbandje, M., Keller, W., Smith, K., Wu, H., Luo, M., Smith, T. J., Rossmann, M. G., Compans, R. W. & Parrish, C. R. (1991). *Science*, **251**, 1456–1464.

Tsao, J., Chapman, M. S. & Rossmann, M. G. (1992). *Acta Cryst.* A**48**, 293–301.

Tsao, J., Chapman, M. S., Wu, H., Agbandje, M., Keller, W. & Rossmann, M. G. (1992). *Acta Cryst.* B**48**, 75–88.

Välegard, K., Liljas, L., Fridborg, K. & Unge, T. (1991). *Acta Cryst.* B**47**, 949–960.

Välegard, K., Murray, J. B., Stockley, P. G., Stonehouse, N. J. & Liljas, L. (1994). *Nature (London)*, **371**, 623–626.

Wu, H., Keller, W. & Rossmann, M. G. (1993). *Acta Cryst.* D**49**, 572–579.

Wu, H. & Rossmann, M. G. (1993). *J. Mol. Biol.* **233**, 231–244.