CHAPTER 4

EXPANSION MODELS

With two good isomorphous heavy atom derivatives available, the obvious direction in which to proceed towards solving the structure is the double isomorphous replacement (DIR) method. However, the prospect is rather daunting for three reasons.

i) The virus has 90 subunits of protein in the crystallographic asymmetric unit. Therefore, unless there is a fluke conformation-dependent binding of one of the metal compounds, there will be at least 90 heavy atom sites. The solution is, of course, constrained by the non-crystallographic symmetry of the virus, but since the orientation relative to the crystal is unknown, this constraint cannot be applied at the outset. Instead, the 90 sites must be found independently and constrained afterwards.

ii) The solution of the heavy atom problem with only 8 Ångstrom data available is by no means guaranteed.

iii) Half of the native dataset took 9 months to collect and process. The prospect of repeating this twice more was not welcomed.

Instead, it was hoped that some starting information could be gained by comparing the diffraction pattern expected from a model structure with that observed. In this way, a set of

phases somewhat better than random should be obtainable, at least at low resolution. Thereafter, the powerful and proven methods of non-crystallographic symmetry averaging (Bricogne, 1974, 1976) could be used to refine the phases reliably (see chapter 5). The expanded TBSV case is extremely favourable because of the 30-fold redundancy of the structure within the crystallographic asymmetric unit.

4.1 Model Description.

As was described in chapter 1, the expansion of the virus is reversible, and the transition between states takes place relatively fast on the time scale of physical chemical experiments. With the knowledge we have of protein denaturation kinetics (Baldwin, 1975), we do not expect that the expansion process involves significant reconformation of the polypeptide chain in each subunit, but instead quaternary structural changes between the subunits, implying relative domain reorientation. Because of the known variability in the interdomain and intersubunit contacts arising from the quasi-symmetry, it is reasonable to expect further variation in these regions rather than within the domains themselves. This is the first assumption of the expansion model.

The second assumption is that the expanded particle retains the full icosahedral symmetry of the compact form, and as much as possible of the quasi-symmetry.

4.1.1 Description of the Contacts between Domains.

Detailed analysis of the many subunit-subunit contacts in compact TBSV suggests the method in which the particle must expand, subject to the first assumption that there is no 'melting' of the secondary structure. The nature of these contacts has been studied by Harrison (1980) with preliminary knowledge of the amino acid sequence. The nomenclature used for a contact is X_{YZ}, where X represents the axis about which the contact is clustered, that is hexamer (H), pentamer (P), trimer (T) or dimer (D). Y and Z refer to the participating subunit positions (A, B or C) in anticlockwise order, looking upon the particle from the outside. Thus the H_{BC} contact is the one that contains the ordered arm attached to the C-position subunit. The various contacts are marked in figure 1.2. It is the spacing of the S-domains that determines the overall size of the particle, and a change in this spacing must therefore be postulated to explain the 10% increase of radius of the particle during expansion and the corresponding 10% increase in linear dimension of the trimer assembly.

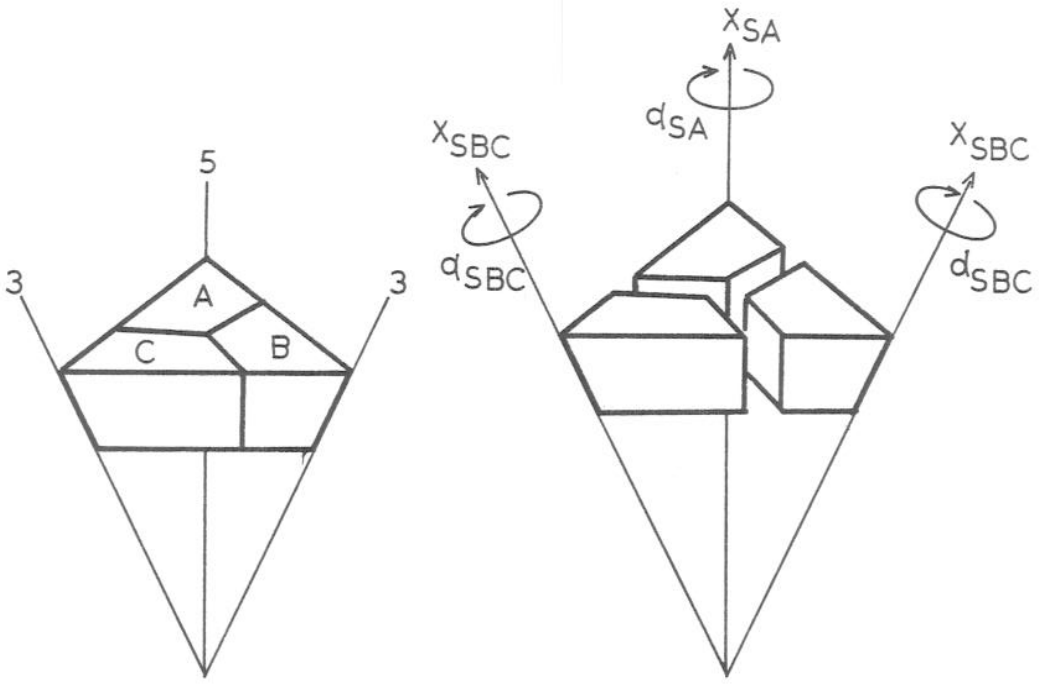
The P-domain dimer contacts consist of a 6-strand plus 4-strand 'sandwich' of β sheets. This is a structure of comparable tensile strength to the interior of protein domains, so the P-domain dimer is more like a single domain than an aggregated state. Indeed, TCV coat protein is known to exist as dimers in solution when the ionic strength is

high (Golden and Harrison, 1981). Whilst this contact is very unlikely to be disrupted, the P-S contact regions are very weak, and could easily be disrupted. Because the hinge extends over several amino acid residues, it is reasonable to expect that new hinge conformations would allow the S-domains to rearrange independently of the P-domains (within certain limits) during expansion, so it is necessary to consider only the S-domains in determining a model. The behaviour of the P-domains can then be predicted by consideration of the quasi-symmetry.

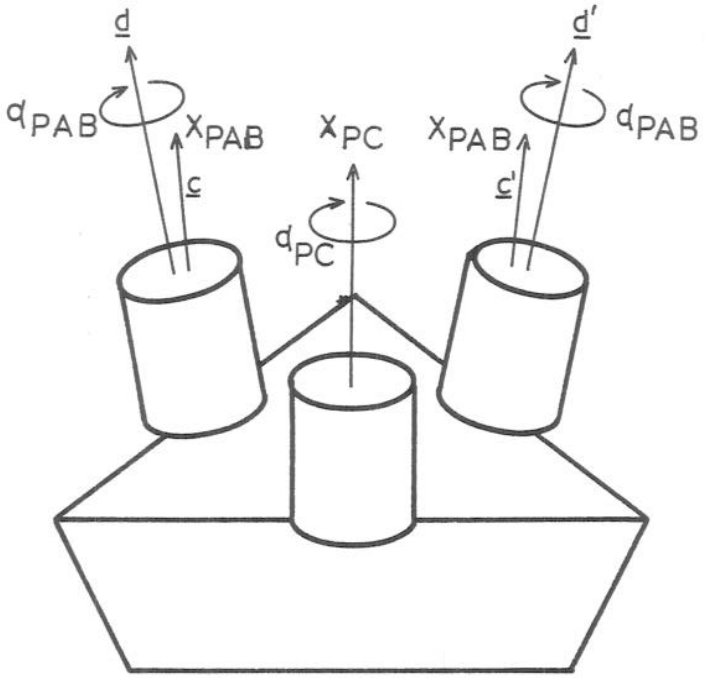
There are basically three conceivable ways that the S-domains can rearrange which preserve the quasi-symmetry of the trimer contacts and the integrity of all the domains including the beta annulus.

i) Disruption of the trimer contacts T_{AC}, T_{CB} and T_{BA} and the dimer contacts D_{AB} and D_{CC}, with retention of all pentamer (P_{AA}) and hexamer (H_{BC} and H_{CB}) contacts. Some rotation about the 5- and 3-fold axes could form a new kind of trimer or dimer contact to stabilise the structure. This is shown in figure 4.1.

ii) Retention of the trimer contacts but disruption of the hexamer/pentamer and dimer contacts. There are two subvariations on this theme. The beta annulus and arm can be preserved if the C position subunits keep their spatial relationship to the particle 3-fold axes; none of the pentamer or hexamer contacts are preserved in this case. The pentamer, hexamer H_{CB} and dimer D_{AB}



(a) S-domains.



(b) P-domains.

Figure 4.1. Schematic diagram of the model for the expansion of TBSV. See table 4.1 for the definitions of the parameters.

contacts could all be preserved if the expansion disrupted the H_{BC} and D_{CC} contacts only; the expanded particle would then consist of twelve solid caps tethered together by the beta annuli and P-domain CC dimer contacts. The arm would have to dissociate from its C position subunit and adopt a new configuration. There is no way for the H_{BC} and all the trimer contacts to be preserved without at least partial melting of the beta annulus.

iii) Retention of dimer contacts only.

The last possibility can be ruled out immediately as one in which all of the subunits do not remain interconnected by their contacts, that is unless some entirely new kind of contact between the S-domains could be utilised. Two pieces of evidence dictate that the first of these is the only reasonable model.

The expansion of TBSV is triggered by the removal of bound calcium atoms from one or both of the sites of the compact virus. The greatest structural change is therefore expected in the vicinity of these sites, which have been located in the middle of the trimer contacts (see figure 1.2) by examination of a difference electron density map calculated for data collected from crystals soaked in EDTA at low pH (Hogle and Harrison, 1981). The Ca^{++} pair is liganded by five aspartate side chains, two from one side and three from the other side of the trimer interface. When the Ca^{++} are removed at high pH, there will be a large local

concentration of negative charge, which will actively force apart the trimer contact.

The relative strengths of the S-domain contacts can be estimated from their relative areas of interaction. A better estimate could be made by counting salt bridges, hydrogen bonds and van der Waals ('hydrophobic') contacts between the participating domains at every interface, but this is difficult without the exact chemical sequence. A disrupted contact has exposed ligands solvated with water molecules: hydrogen bonds would be expected to be more or less indifferent to the state of the contact, but salt bridges (in solutions of low ionic strength) and van der Waals contacts would be more sensitive. Examination of the details of the various contacts in compact TBSV (Harrison, 1980) suggests the following ordering of contact strengths:

$$H_{BC} > H_{CB}, P_{AA} > T > D.$$

The hexamer/pentamer contacts are by far the most extensive, each having at least three discernable salt bridges and at least five hydrophobic residues in van der Waals contact in a scheme of alternating stripes of polar and hydrophobic patches. Such a scheme may be important in the alignment of the subunits during self-assembly. The H_{BC} contact is quite different from the H_{CB} and P_{AA} (which are similar to each other); the relative orientations of the two S-domains changes by approximately 20^{o} about an axis lying along the interface and parallel to the viral surface (the 'fulcrum'

of Harrison, 1980). The cleft that forms beneath the H_{BC} contact because of this conformation difference is occupied by the arm of the C subunit, which interacts by means of additional salt bridges, extending the area of the contact and rendering it the strongest of the three. The trimer contacts are stabilised largely by the bound Ca^{++} ions in the compact virus, and partly by van der Waals contacts around the quasi 3-fold axis. Even with the Ca^{++} bound, the contacts are not as extensive as the hexamer/pentamer ones. The dimer contacts are smaller still, and mainly hydrophobic.

The conclusion of this comparison of contact strengths is that the least likely contact to be disrupted by the expansion is the H_{BC} one; this fact alone is consistent only with the first proposed model above.

Thus we expect the S-domains to retain their spatial orientation to the 5- and 3-fold axes during expansion in order to conserve the hexamer and pentamer contacts. This is fairly restrictive and so limits the number of degrees of freedom in the model. The P-domains, which were not included in the arguement above, form the essential links between the pentamer and hexamer clusters to maintain the integrity of the expanded virus. Because the underlying S-domain contacts are not conserved, new hinge configurations will be needed to permit this. The only constraints that can be applied to the P-domains are the conservation of

their own dimer contacts (far more reasonable than the alternative that these are broken in favour of conservation of the S-P contacts) and the particle 2-fold symmetry which applies to the CC P-domain dimer only.

4.1.2 Simplified Starting Model.

With this third assumption about the expansion process, that the S-domain pentamer/hexamer contacts and P-domain dimer contacts are conserved, we have a model with 12 free parameters. These are listed in table 4.1 with the names that will be used henceforth to refer to them. Spherical polar coordinates are used to describe the 6 degrees of freedom of the general rigid body transformation of the AB P-domain dimer, which is unconstrained by the virus symmetry. This parameterisation is better than a straightforward rotation plus translation because it is more amenable to the imposition of constraints to limit the number of degrees of freedom in the model. It is still an orthonormal description, so does not introduce unnecessary correlations between the parameters. It is expected that the direction of the axis of the AB P-domain dimer rotations will be close to the local diad axis that relates the A subunit to the B one, but this is not a good approximation for the direction of the radial translation, which will more likely bisect the directions of the point group 3-fold and 5-fold axes.

x_{SA}	Expansion displacement of A position S-domains along particle 5-fold axis.
x_{SBC}	Expansion displacement of B and C position S-domains along particle 3-fold axis.
x_{PAB}	Expansion displacement of A and B position P-domains along direction $\underline{c}$ defined below.
x_{PC}	Expansion displacement of C position P-domains along particle 2-fold axis.
α_{SA}	Clockwise rotation angle of A position S-domains about particle 5-fold axis.
α_{SBC}	Clockwise rotation angle of B and C position S-domains about particle 3-fold axis.
α_{PAB}	Clockwise rotation angle of A and B position P-domains about direction $\underline{d}$ defined below.
α_{PC}	Clockwise rotation angle of C position P-domains about particle 2-fold axis.
(c_x, c_y, c_z)	Direction cosines for expansion displacement direction of AB P-domain dimer.
(d_x, d_y, d_z)	Direction cosines for direction of rotation of AB P-domain dimer.

Table 4.1. Definitions of the 12 parameters describing the most general model of the expansion of TBSV. The null model with all parameters zero describes the positions of the domains in the compact virus. A diagram of this model description is shown in figure 4.1.

Thus, the most general description of the expanded virus crystals, subject to the three assumptions stated, is a set of 13 parameters, being the 12 parameters of table 4.1 plus the crystallographic packing angle, Θ, referred to in chapter 2. The model is that of figure 4.1. It is infeasible to conduct a search for 13 independent parameters without some a priori information about the answer, so we must begin by restricting the model to a smaller dimension space, and generalise at a later stage. The crudest conceivable model that retains expansion features is described by 2 parameters:

Θ free

$x_{SA} = x_{SBC} = x_{PAB} = x_{PC} = x$ free

$\alpha_{SA} = \alpha_{SBC} = \alpha_{PAB} = \alpha_{PC} = 0$ fixed

Direction of AB translation: fixed

Direction of AB rotation: irrelevant.

The rotation angles for the S-domains must be very close to zero to prevent these colliding with each other; the P-domains could be at any angle, but these have approximate cylindrical symmetry so will not look very different at low resolution. The direction of the AB P-domain expansion displacement is fixed here to be half way between the 3-fold and 5-fold axes, and the expansion distances are the same for all subunits (each in their appropriate direction). This model is now in a practical form for an R-factor search.

4.1.3 Proposed Strategy for Obtaining Phases.

The method proposed for deriving phases for the structure is to use this very simple model to find optimal values for x and Θ and then gradually introduce more and more parameters to improve the fit. After this, it is hoped that non-crystallographic symmetry averaging will be capable of refining the phases calculated from the best model to a fairly accurate final set. The most important question to be raised in criticism of this proposed scheme is one of bias: if a model, particularly one that itself has icosahedral symmetry, is used to provide a starting phase set for symmetry averaging refinement, will that not bias the final structure to look like the model? It is known that a 'map' of a structure calculated with random amplitudes and correct phases largely resembles the original structure. In order to answer this question, we must perform such experiments as: if part of the model is removed at the start, does it reappear after refinement? If additional features are added, do they diminish in strength as the refinement proceeds? We will return to the bias question after looking at some results.

The heavy atom derivatives are also very useful tools for comfirming the phases. If a heavy atom difference map, phased in the way described above, shows punctual features resembling atoms, then it suggests that the phases are meaningful. Moreover, if the positions of those sites

display the symmetry and quasi-symmetry of the virus, then the suggestion is strongly supported. Since the metal-containing reagents used to make the derivatives are similar to those used for the compact virus structure, there is even a good chance that the same sites will be hit and a comparison between the observed and expected positions will be possible. Since the phases are derived from observations and a model alone, no information about the compact virus heavy atom positions is ever included, so this last test is a completely independent control.

4.2 Calculation of Structure Factors.

There are two basic ways of calculating structure factors for a model based on a related structure.

i) Density corresponding to the individual domains can be carved out of an electron density map and, by suitable skewing and interpolating, be reconstructed in a new location. The carving operation is a fairly difficult one because great care must be taken not to introduce sharp steps at the boundaries which would lead to serious distortion of the calculated structure factors. To compensate for this, it would presumably be necessary to calculate a map of much higher resolution than the required structure factors. The other serious drawback with this method is that all the noise and heavy atom artefacts of the starting map are included too.

ii) Structure factors can be calculated directly from a set of atomic coordinates by direct summation of the trigonometric series. However, the number of atoms in the asymmetric unit is about one million. Even to evaluate a few hundred structure factors would be prohibitively time consuming, and would leave a lot to be desired in the reliability of the comparison with observations; a few thousand structure factors would be much safer. Some time could be saved by using isotropic form factors for entire amino acid residues, which might be a good enough approximation at low resolution only. The sum would be performed over the 10^5 residues in the asymmetric unit, but introduces errors of an unknown magnitude.

Instead it was decided to take advantage of the speed of the first method, which employs an FFT (Cooley and Tukey, 1965), and the accuracy of the second, which is based on an idealised, stereochemically exact version of the initial structure. To do this, a 'fake' map of the model structure is first calculated from atomic coordinates transformed according to the current model parameters, using an assumed density distribution for each atom. This is Fourier transformed to produce the desired structure factors. As was suggested above, the number of atoms can be reduced by replacing all of the atoms of a residue by a single atom of appropriate radius, when only low resolution structure factors are required.

The problem of generation of a map from atomic coordinates has been tackled by Agarwal (1978) and others as part of a 'fast Fourier' method of reciprocal space refinement. The method places the density for each atom as a distribution which is the sum of three 3-dimensional Gaussian functions. No attention is paid to aliasing; he merely suggests that the map should be generated on a grid with less than one third of the spacing of the highest resolution structure factors desired. For the application of performing an R-factor search for expanded TBSV, it is anticipated that a fairly large number of fake map generations will be required, so the efficiency is of fundamental importance: an 8 Ångstrom FFT for this structure takes 7 minutes on the department VAX 11/780; at 5.5 Ångstroms, this would take 25 minutes, which is unacceptable. The first problem to be solved therefore is how to get rid of the aliases at the stage of map generation, so that a minimally sampled FFT can be used to calculate structure factors.

4.2.1 The Aliasing Problem and its Solution.

The problem of aliasing is demonstrated in figure 4.2. It arises when a density function is sampled at a frequency lower than its Shannon frequency. The structure factors calculated <u>do</u> correspond exactly to the density function at the sample points (because the discrete Fourier transform is exact), but not elsewhere. Evaluation of the Shannon interpolation formula (1949) between the sample

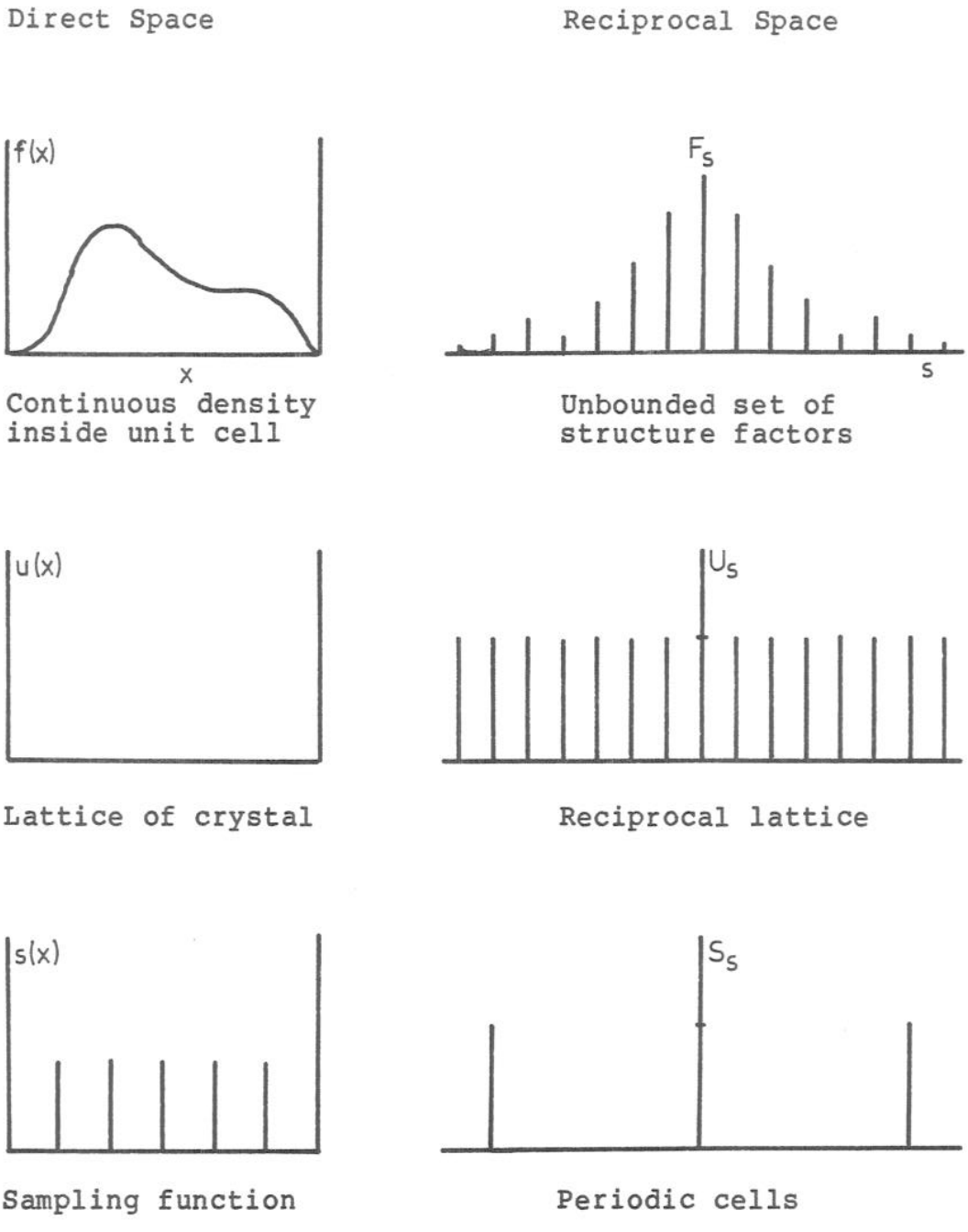


Figure 4.2. One dimensional demonstration of the origin of aliases with the discrete Fourier transform. Each vertical bar represents a delta function enclosing an area equal to the height of the bar. Continued on next page.

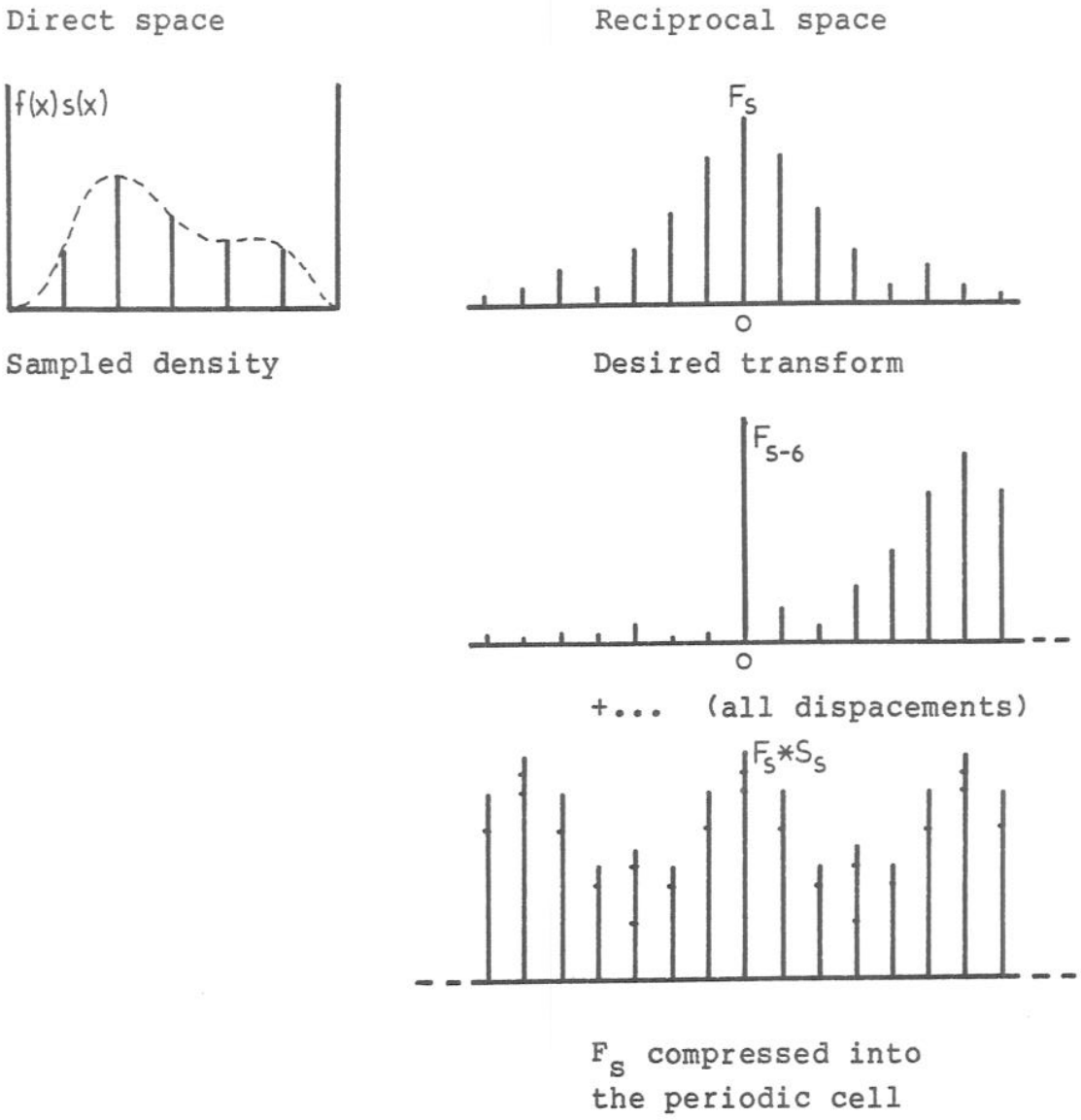


Figure 4.2 (continued). One dimensional demonstration of the origin of aliases with the discrete Fourier transform. When the density is sampled at only a finite number of points, the resulting structure factors are compressed into a periodically repeating cell of reciprocal space. The final discrete Fourier transform is a sum of the analytic Fourier transform with <u>all</u> origin displacements. When the latter has frequency components above the sampling frequency, these contribute aliases to the fundamental frequency band.

points does not yield the same density values at all points, but only where the density is sampled. The problem is that the structure factors calculated by discrete Fourier transformation of the sampled density deviate from the first few terms of the exact Fourier series that would correspond to the measured values (figure 4.2(a)), because of contamination by Fourier terms from above the sample frequency. When the density function is sampled more frequently than the Shannon frequency, there are no such terms to contribute an alias. The problem is worse in three dimensions because there are six adjacent frequency bands contributing aliases instead of two.

The exact solution to this problem is shown in figure 4.3. The density is prefiltered by convolution with a sin(x)/x function which removes all frequencies greater than the sampling frequency. The Fourier transform of sin(x)/x is a box function, so such convolution is the real space achievement of perfect low-pass filtering. The discrete Fourier transform is then exactly equal to the first few terms of the complete Fourier series. The application for calculating exactly the first few terms of a Fourier expansion of a set of point atoms is obvious: each atom is replaced by a sin(x)sin(y)sin(z)/xyz function (convolution of sin(x)sin(y)sin(z)/xyz with a delta function) and the discrete FFT is calculated. One approximation is required though for practical applications. A cutoff distance must be imposed on the function, or else every atom

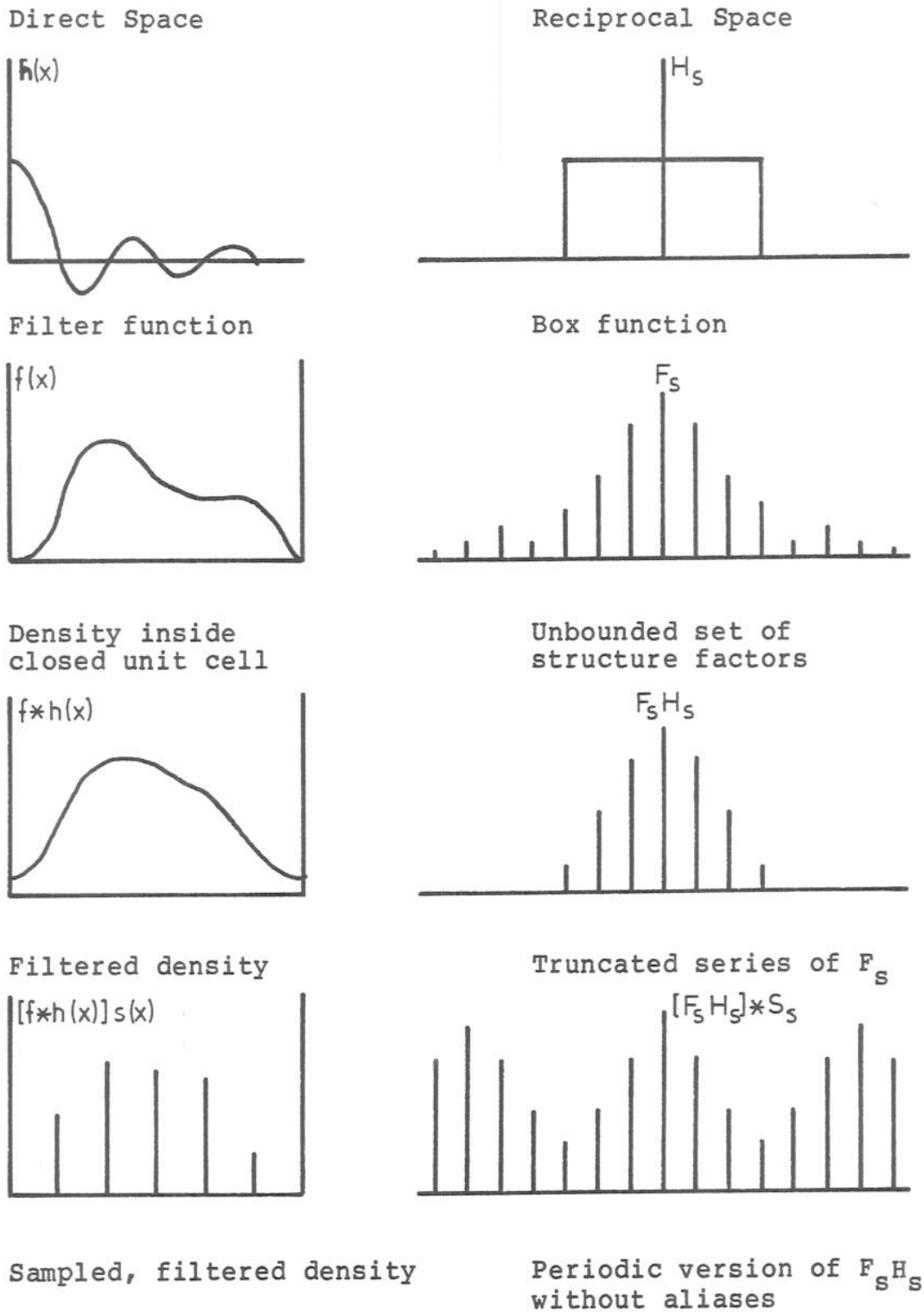


Figure 4.3. Removal of all aliases by means of a sin(x)/x filter.

would make a contribution to every point in the map; since the function decays only as 1/x along the map axes, a fairly large number of contributions are needed for accuracy.

4.2.2 Application to Map Generation.

In the current application we wish to calculate a map from pseudo atoms that each represent whole residues, so are not point-like, but which have a size comparable to the grid spacing at 8 Ångstroms. If we represent each pseudo atom as a spherical Gaussian function, then we must convolve this with the sin(x)/x function to obtain the filtered contribution. There is no analytic expression for this convolution, but we can make use of its expansion as a double power series:

$$g(x) = \left[\frac{2}{\pi\sigma^2}\right]^{1/2} e^{-x^2/2\sigma^2}$$

$$h(x) = \frac{\sin(bx)}{bx}$$

$$\Rightarrow \quad g*h(x) = \frac{b}{\pi} \sum_{i=0}^{\infty} \frac{u^i}{2i+1} \sum_{k=0}^{i} \frac{v^k}{(2k)!\ (i-k)!} \qquad (4.1)$$

$$\text{where} \quad u = \frac{-\sigma^2 b^2}{2} \qquad v = \frac{2x^2}{\sigma^2}$$

For purposes of testing, a computer program was written to calculate a 6 Ångstrom fake map of compact TBSV from single residue pseudo atoms using this formula. Structure factors

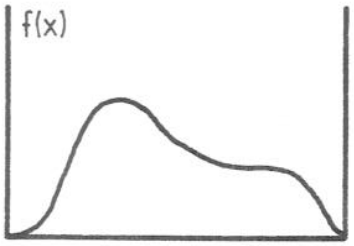
from this map calculated to the Shannon limiting sphere agreed with the observed data with an R-factor of 0.45 (see below for definition) and a mean phase difference of 50°. Even with the evaluation of the power series performed by a subroutine written in VAX assembler language whereby the recursive floating point multiply/accumulate instruction 'POLYF' can be utilised, the program took 50 minutes of CPU time to execute, so was clearly inadequate, and was subsequently abandoned.

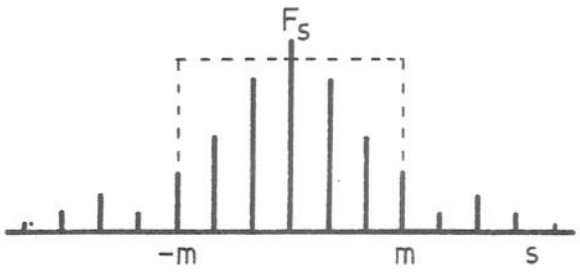
4.2.3 Gaussian Filter Method.

Instead of attempting to filter the aliases perfectly, it was decided to make do with a Gaussian low pass filter which is much easier to apply computationally. Here the density function is convolved with a Gaussian of appropriate width (see figure 4.4); this means the calculation is especially easy for spherical Gaussian pseudo atoms as the convolution of two Gaussians is itself a Gaussian. Figure 4.4 shows the determination of the optimal filter function: it is the ratio of the function value at the outermost wanted resolution to its value corresponding to the innermost alias that must be maximised. In the infinitesimal limit, this means that the slope of the filter function is greatest at the cutoff resolution. For a Gaussian, this occurs when $x = \sigma$, the standard deviation of the distribution.

Direct Space

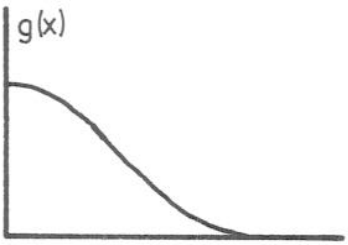
Reciprocal Space

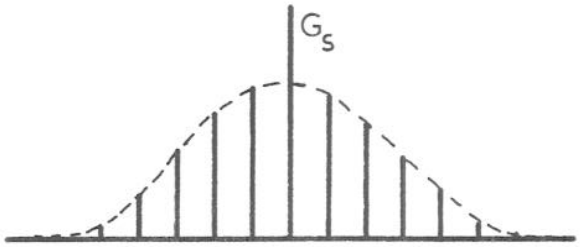




Density, f(x), inside closed unit cell

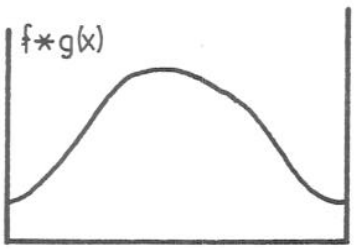
Unbounded set of structure factors, F_s

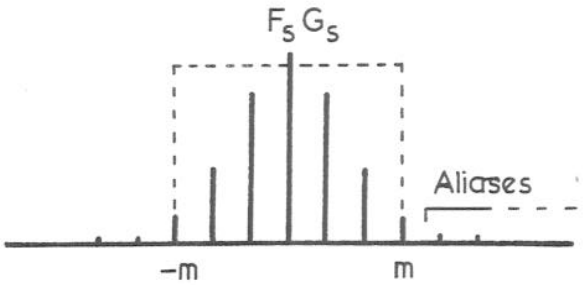




Convoluted with Gaussian filter function, g(x)

Each multiplied by G_s, transform of g(x)





Filtered density, f*g(x)

Filtered structure factors, $F_s G_s$

Figure 4.4. Gaussian filtering. Continued on next page.

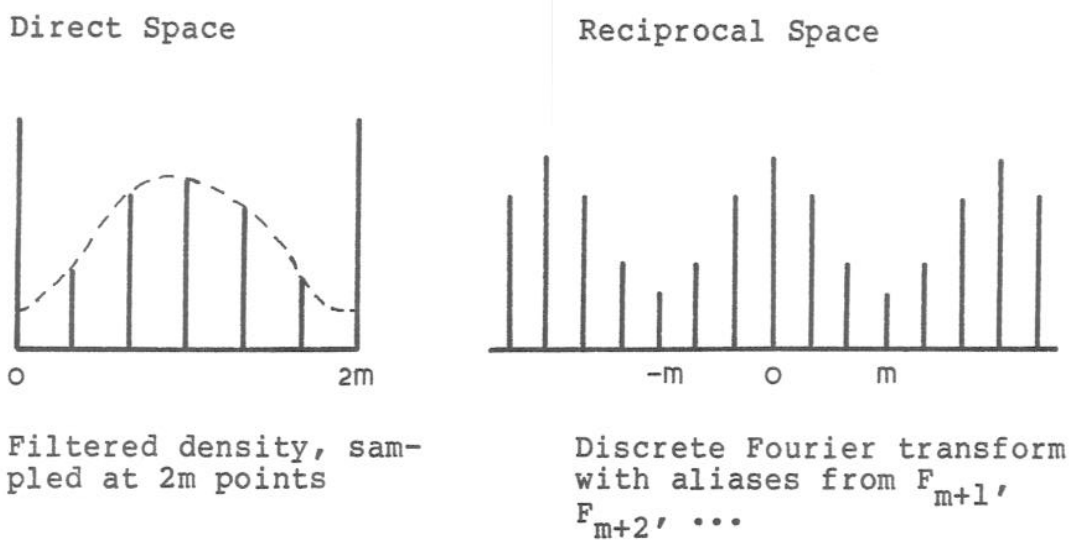


Figure 4.4 (continued). Gaussian filtering, analogous to the perfect filtering of figure 4.3. When there are 2m sample points, the outermost wanted structure factor is F_m. The innermost aliases to be rejected arise from F_{m+1} adding to F_{-m}, F_{-m-1} adding to F_m, etc. The optimal width of the Gaussian, g(x), is the one that gives the largest ratio of G_m to G_{m+1}. I.e. the slope of the reciprocal Gaussian, G_s, should be greatest at s = m.

The overall effect of filtering with this function is that an 'artificial temperature factor' is applied to the structure factors. This is of no importance as it is taken into account in the scaling of observed to calculated structure factors at the stage of comparison to produce the R-factor.

The generation of the fake map by this method is straightforward. The width of each pseudo atom (see below) is combined with the optimal filter function width, which is equal to twice the grid spacing of the map. The convolution width is the square root of the sum of the squares. Then to each map grid point within 2 or 3 Gaussian widths of that atomic position is contributed a density from the spherical Gaussian distribution. The choice of radial cutoff level (2 or 3σ) is explained below.

4.2.4 Atomic Coordinates.

The set of atomic coordinates was obtained from the stereochemically justified coordinates of the C position S- and P-domains (Harrison, et. al., 1978). These were measured directly from a Kendrew model built inside a Richards optical comparator (1968). The coordinates of a number of corresponding alpha carbon positions in the A and B position subunits were measured directly and the optimal least squares superposition transformations from the C position were calculated (separate transformations for the S-

and P-domains). Three residues in the hinge region were omitted. The coordinate lists were reduced to pseudo atoms by calculating the centroid and the radius of gyration of the atomic positions for each residue. These values define the position and radius of each pseudo atom. A computer program was written to read through these two lists of pseudo atom coordinates for the S- and P-domains each three times to generate the full 'trimer' (viral asymmetric unit) by application of the superposition transformations. At the same time the additional transformations of the expansion model were applied, with the expansion parameters taken from the input to the program. Tests of the program for the compact structure could be readily achieved by running the program with null parameters. The 36 residues of the beta annulus were removed in the A and B position subunits.

Because it was anticipated that the fake map generation, FFT and structure factor comparison programs would need to be run fairly often, a great deal of time was spent optimising the performance of these. The noteworthy of these optimising features are mentioned.

4.2.5 Organisation of Programs.

To save input/output operations between steps of the procedure, the total number of separate programs was reduced to two: the various steps of each were converted into subroutines called by the previous step. The passing

of subroutine parameters is much faster than the corresponding file handling operations. The first program performed the icosahedral expansion of the coordinates, the generation of the Gaussian distributed density contributions, the sorting of these into the various sections of the fake map, and the reconstitution of the map which was then written out. The second program was a modification of the direct access file version of Ten Eyck's FFT structure factor program for space group $P2_1$ (Ten Eyck, 1973) which inverted this map. $P2_1$ is a subgroup of C2 but has no systematic absences, so the program was modified to skip over the calculation of the absent reflections during the third pass of the FFT. Rather than write out the calculated structure factors, it read in the corresponding observed values to be compared, calculated the scale factor between them based on the first 1,000 observations in each resolution range, and then calculated R-factors and correlation coefficients for each range. No attempt was made to apply proper resolution-dependent (i.e. temperature factor) scaling to the calculated structure factors; instead, a number of independent resolution ranges were scaled separately.

Additional time savings were made by interpolating from a look-up table to obtain the values of the exponential function in evaluating the Gaussian distributed density.

4.2.6 Core Sort.

An enormous saving of time was made by the use of a 'bin' sort for the density contributions. The VAX is a virtual address machine which means that very large areas of storage can be accessed directly by a program; indeed an entire map could be addressed and filled directly with density contributions as the psuedo atom list is read. However, such a large amount of physical memory is not available, so that this strategy would cause the machine to swap pages of virtual memory into and out of the limited physical region at an enormous rate, drastically reducing the speed of execution. The alternative strategy is to write out one long list of all the density contributions, together with an address in the final map to which they belong. This list is sorted on the section number (the y component of the map address) and then read back in to fill up one section of the map at a time. Random access within a single section is possible with the available physical memory allocation.

This second method, on the other hand, requires two extra sets of input/output operations and an additional sort step. The VAX/VMS system software does not offer an efficient sort routine, so an alternative bin sort was devised. The requirements for this kind of sort are rather special: the range of possible values of the sort variable is known and the number of different values is less than 100. Moreover, the number of records with each possible

value is approximately the same, let us say 10,000 records. Therefore an array of dimension 10,000 x 100 can hold all of the contributions; since this array is addressed sequentially in its long direction, both during reading and writing, the active region can always be held by 100 pages of physical memory (provided the array is dimensioned the right way around!). The application is straightforward: each density is packed along with its (x, z) address into 4 bytes and written into the next available column of the y'th row of the array. For reconstitution of the map, the array is read row by row and a section is filled up and written out for each. The time to generate an 8 Å ngstrom TBSV fake map was reduced to less than 5 CPU minutes from 30 CPU minutes using the VAX sort routine. The principle has since been very sucessfully generalised by C. Steele (unpublished) into a freestanding sort routine 'BSORT' for use with the skewplanes and density averaging procedures (see chapter 5).

4.2.7 Tests with Compact TBSV.

To test the programs, it was decided to use the unexpanded model in the space group of the compact structure (I23, a = 383.2Å) and compare with the observed compact virus dataset. The crystals of the compact structure are body centred and so this would have meant making additional modifications to the fake map program which packs the particles in a face centred fashion. Rather than risk the possiblity of the programs working for the orthogonal situation

but not the monoclinic one, the compact virus dataset was reindexed as space group C2 ($a = 2^{1/2} \times 383.2$, $b = c = 383.2$, $\beta = 135^{o}$) which is a subgroup of I23 and is face centred (see figure 2.3). The index transformation required was:

$$h' = -l \pm h$$
$$k' = k$$
$$l' = l$$

with both signs as well as all three permutations of the compact (h, k, l) included. After sorting, these new indices could be used with the fake map programs without any alteration and with all the cell constants fairly close to their values for the crystals of the expanded structure. This is therefore a very critical test of the correct functioning of the whole model system.

The statistics for these trial runs are given in table 4.2. Two variations of parameters were tried: the Gaussian cutoff for each pseudo atom was tried at both 2 and 3 standard deviations. The latter is clearly an improvement and does not cost much extra time. A greater suprise was the discovery that the use of half the number of pseudo atoms (by replacing adjacent pairs by a single atom of suitably larger radius) leads to a <u>smaller</u> R-factor whilst saving a great deal of time. This last result is unexplained but gratefully accepted! The time taken for the FFT and comparison step was 7 minutes in all cases, making a total of 11.5 minutes per search point for the whole procedure in

(a) R-Factors for 16 to 9 Ångstrom Data.

	Every Residue	Pairs of Residues
Number of atoms	882	441
2σ cutoff		0.372
3σ cutoff	0.368	0.361

(b) Execution Times Taken (Minutes and Seconds).

	Every Residue	Pairs of Residues
2σ cutoff		3:23
3σ cutoff	6:58	4:35

Table 4.2. Optimisation of fake map generation method tested with compact TBSV. Variations of the Gaussian cutoff distance in standard deviations and the number of pseudo atoms in the model (one per amino acid residue or one per two residues) are investigated.

its final form at 9 Ångstroms resolution for the compact structure, or 13 minutes for the expanded form to 8 Ångstroms. Thus an entire scan of a single parameter could be performed in an hour, which is a very reasonable amount of time for a computer experiment.

The final R-factor of 0.36 is an indication of the level of approximations made by the described procedure and inherent in the atomic coordinates used. This value sets a lower level target for the R-factor search for the expanded structure. The expected R-factor for a random structure unrelated to the observed structure factors but scaled correctly is $2 - 2^{1/2} = 0.59$ for non-centric reflections (see appendix A).

4.3 R-Factor Searches.

The crystallographic R-factor for the comparison of a set of n observed structure factors, $\{F_i^{obs}\}$, and the corresponding $\{F_i^{calc}\}$ is defined

$$R = \frac{\sum_{i=1}^{n} |F_i^{obs} - F_i^{calc}|}{\sum_{i=1}^{n} |F_i^{obs}|}$$

The statistical correlation coefficient is not widely used in crystallography because it is insensitive to small differences and harder to calculate. It is defined

$$r = \frac{n \sum_{i=1}^{n} F_i^{obs} F_i^{calc} - \sum_{i=1}^{n} F_i^{obs} \sum_{i=1}^{n} F_i^{calc}}{\left[n \sum_{i=1}^{n} F_i^{obs\,2} - \left(\sum_{i=1}^{n} F_i^{obs}\right)^2\right]^{1/2} \left[n \sum_{i=1}^{n} F_i^{calc\,2} - \left(\sum_{i=1}^{n} F_i^{calc}\right)^2\right]^{1/2}}$$

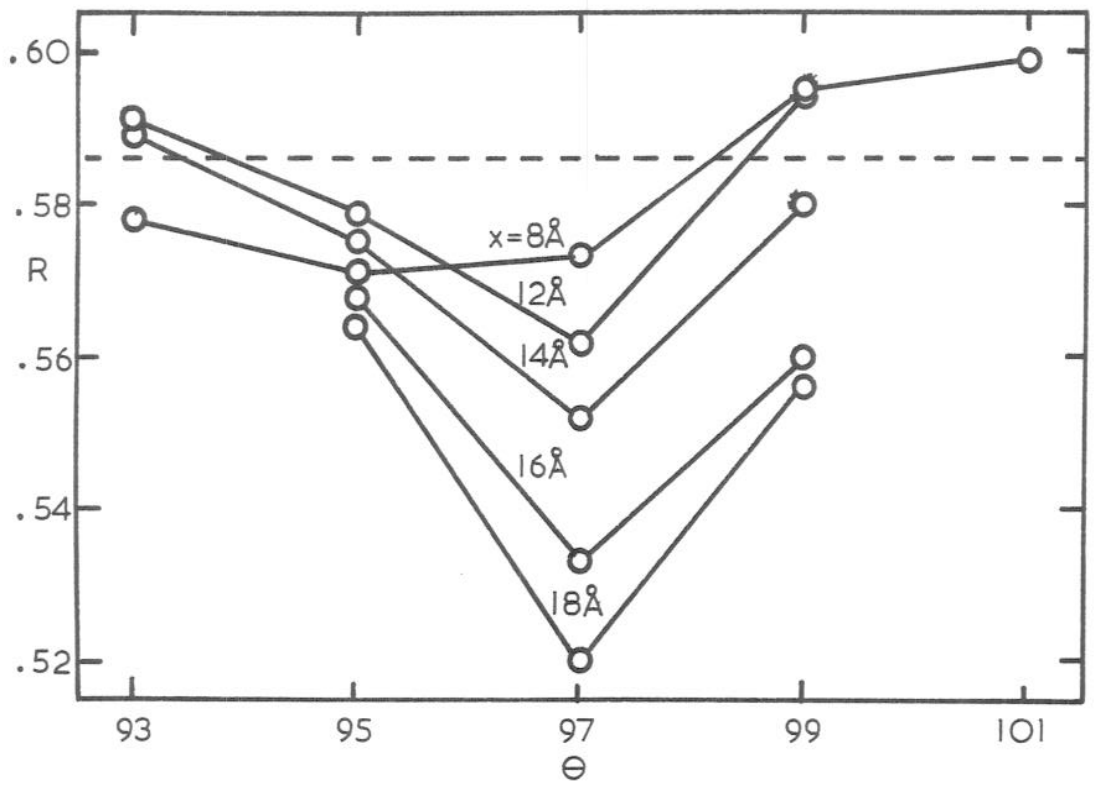
The correlation coefficient has the advantage that it is very sensitive to small correlations, having a value of zero for uncorrelated variables, and that it is completely insensitive to the absolute scale of one variable to the other.

As mentioned above, these two coefficients were calculated for every model investigated and their variation as a function of the model parameters was used to determine the best model, that is the one with the smallest R-factor or largest correlation coefficient.

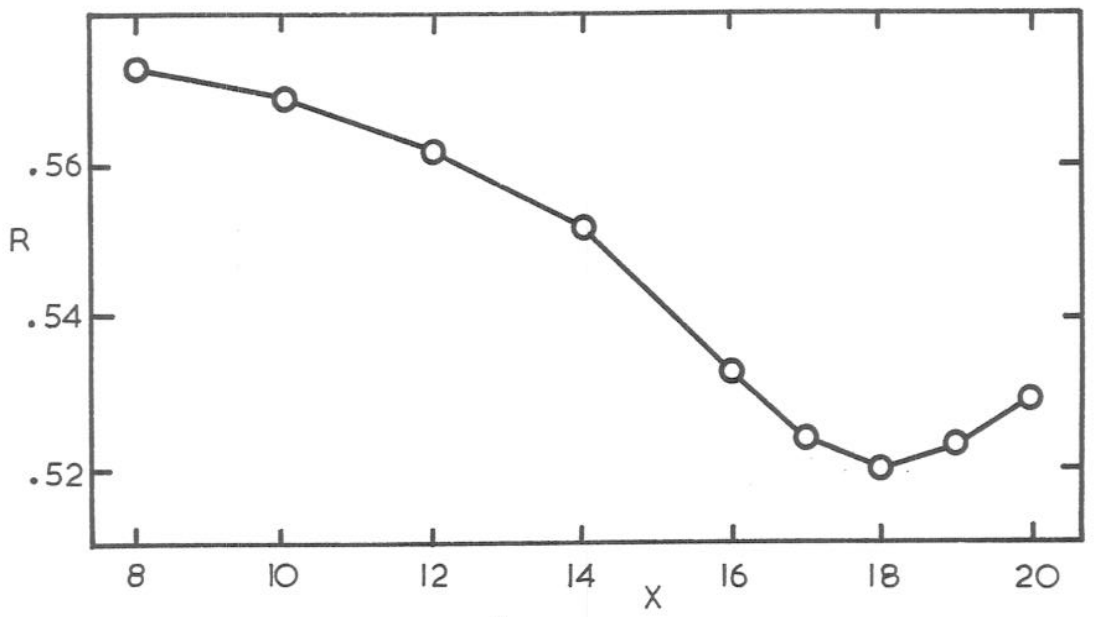
4.3.1 R-Factor Search for Expanded TBSV.

The R-factor at each point of a search with expanded virus data between 35 and 16 Å ngstroms took just 4 CPU minutes to calculate. 3,658 structure factors were compared with their calculated values in four resolution shells. The two parameter model was used at first, starting with $\Theta = 97^{\circ}$, the value derived from the spike positions measured on a b* precession photograph (see chapter 2), and $x = 8$Å, the value predicted by the crystal packing studies (section 2.3.2). Figure 4.5(a) shows the search for Θ with various different values of x. Two conclusions result:

i) The optimal value of Θ is 97°, and this is relatively insensitive to the choice of x, except for the initial



(a) Search for Θ at various values of x. The profile is sharper when x is closer to its best value of 18Å. The R-factor for a random structure is indicated by the dashed line.



(b) Search for x at Θ = 97°.

Figure 4.5. R-factor searches for Θ and x in the 2 parameter model.

value of x = 8Å which is so far off that a false minimum of $\Theta = 95^{o}$ is suggested.

ii) The profile in Θ becomes steeper when the value of x is closer to its best value of 18Å. The search is more sensitive when the answer is close.

When the search is plotted the other way around in figure 4.5(b), the value of x = 18Å is clearly seen to be optimal.

The discrepancy between x = 8Å from the crystal packing and x = 18Å from the R-factor search probably indicates that some modification of the particle at the packing contacts takes place, so that neighbouring particles do not overlap. x = 18Å agrees very well with the values obtained for the radius of the expanded particle in solution (see chapters 1 and 2).

4.3.2 Extension to a 3-Parameter Model.

In an attempt to introduce more degrees of freedom into the model so that better agreement could be obtained, the following model was tried:

Θ free

$x_{SA} = x_{SBC} = x_S$ free

$x_{PAB} = x_{PC} = x_P$ free

$\alpha_{SA} = \alpha_{SBC} = \alpha_{PAB} = \alpha_{PC} = 0$ fixed

Direction of AB translation: fixed

Direction of AB rotation: irrelevant.

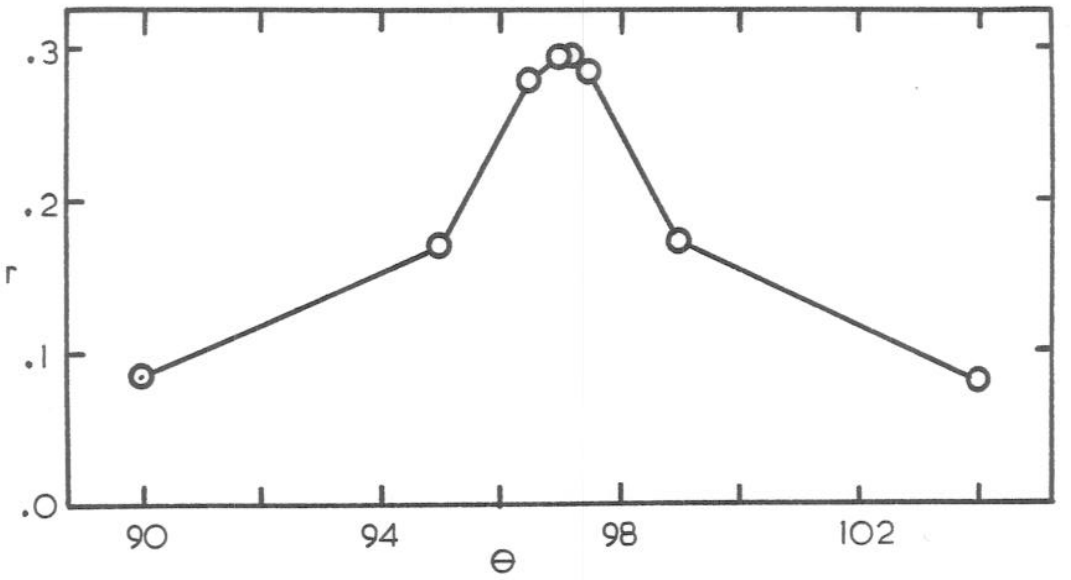
Keeping $\Theta = 97^{\circ}$, a broad search of x_S and x_P was made; the results are displayed in table 4.3. The amount of variation close to the minimum was disappointingly small. The optimal values were $x_S = 18\text{Å}$ and $x_P = 20\text{Å}$, each with an estimated error of $\pm 2\text{Å}$. The shape of the minimum was checked for elongations in directions other than the x_S or x_P that would be indications of correlation between the variables, but nothing significant was found.

It was considered doubtful that any further fine-tuning would give meaningful results, so a final search was made for Θ, this time to 12 Ångstroms resolution using the optimal values above. The results are shown in figure 4.6. The profile now has a highly symmetric shape and the best value of 97.0° has an estimated error of $\pm 0.3^{\circ}$.

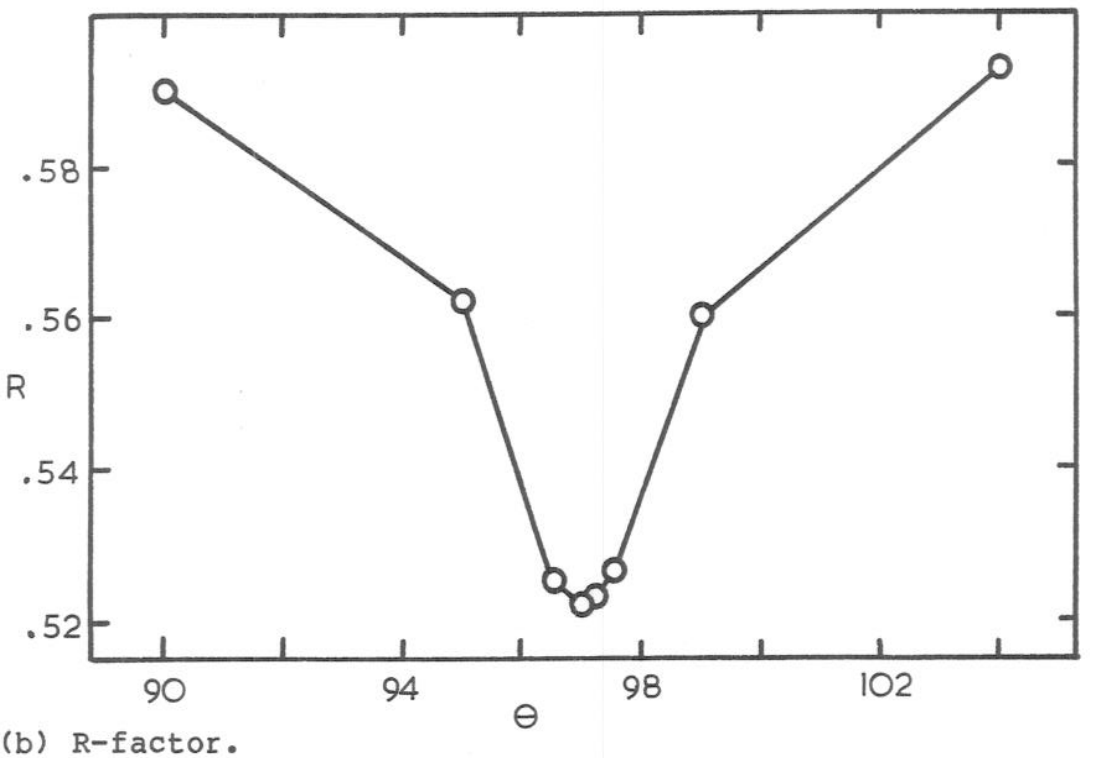
At this point, it was clear that phases could be generated that were better than random, as indicated by the final R-value of 0.52 and correlation coefficient of 0.30, so it was decided to attempt phase refinement, in the hope that more information relating to a better model would be forthcoming.

	x_S				
x_P	16	17	18	19	20
17		.524		.528	
18		.520	.520		
19	.521	.518	.518	.523	
20		.517	.517		.529
21		.520			

Table 4.3. Two dimensional R-factor search for the expansion displacements x_S and x_P. Θ is fixed at 97°; distances are in Angstroms.



(a) Correlation coefficient.



(b) R-factor.

Figure 4.6. Search for the value of Θ in the best 3 parameter model.