INTRODUCTION TO THE USE OF NON-CRYSTALLOGRAPHIC SYMMETRY IN PHASING

MICHAEL S. CHAPMAN
Department of Chemistry & Institute of Molecular Biophysics, Florida State University, Tallahassee, FL 32306-3015, USA.

Abstract: The methodology is reviewed with particular attention to developments since prior review, and to aspects most important to the improvement and extension of crudely approximated initial phases as might be encountered in *ab initio* structure determinations. The fundamental similarity between symmetry and multiple crystal averaging is emphasized.

1. Introduction

Non-crystallographic symmetry (NCS) alone has not yet been used to determine *ab initio* a macromolecular structure. However, it provides one of the most powerful methods for improving approximate phases, and has become an important part of *ab initio*, experimental and molecular replacement phase determinations. Its use required the development of the methods of molecular replacement [1]. By the mid 1970's real-space averaging algorithms conceptually similar to those currently in use had been implemented [2]. Bricogne's paper [2] laid out some of the problems that have been addressed at least partially in the last 20 years. With increasing confidence in the methods, NCS became widely exploited in the 1980's. The methodology has matured rapidly and an excellent review has been published [3]. In the prior review and in other literature, "symmetry averaging" sometimes comes under the umbrella term, "molecular replacement", due to methodological similarities with the use of one structure to determine the phases of another. Here, the focus is "averaging", including the use of non-crystallographic symmetry (NCS) and multiple crystal forms. Molecular replacement is covered elsewhere in this volume.

Prior to the widespread use of NCS, multiple copies of a molecule in an asymmetric unit was considered a substantial disadvantage in structure determination. Likewise, the tendency of proteins to crystallize in different forms was at best an inconvenience. A

S. Fortier (ed.), Direct Methods for Solving Macromolecular Structures, 99–108.

testament to the newly realized power of averaging is the joy in at least the author's laboratory with which redundancy in the same crystal or different crystals is greeted.

Redundancy is the source of the phase information. A diffraction pattern is the continuous molecular transform sampled at discrete lattice points. A set of structure amplitudes may contain only ~ ½ the information required to reconstruct the molecular transform at a given resolution, because the phases are missing. Data from another crystal form samples the same molecular transform at different points. If the resolutions are the same, no detail is added. Rather, with the doubling of the number of samplings, there is, in principle "only one" set of phases to the molecular transform that can give both sets of amplitudes. Incorporation of multiple copies of a molecule within the same crystal has a similar effect. The unit cell size is increased, and more reflections can be measured at a given resolution, but the unique part of the structure remains the same.

For computational expediency, the redundancy is usually exploited with iterative real-space calculations:

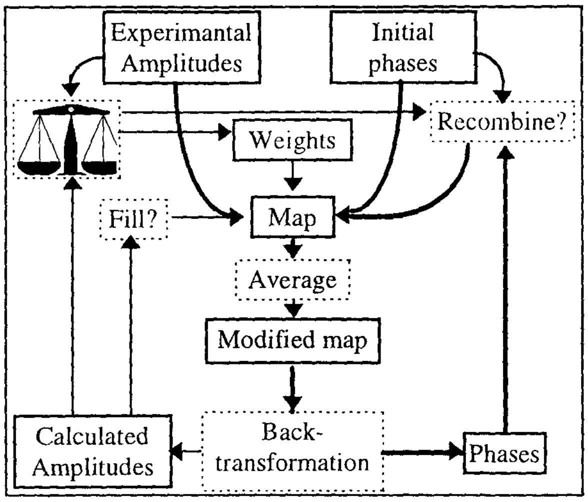

Figure 1: Phase refinement by iterative real-space averaging. The basic procedure follows thick arrows: (a) a map is calculated from the observed amplitudes and initial phases; (b) the map is replaced by one that has been averaged. (Other modifications can also be applied, such as solvent flattening.); (c) Back-transformation yields new phases which can be merged with the initial ones, or combined directly with the observed amplitudes for calculation of the improved map for the next iteration. Common embellishments to the basic mechanism are shown with thin arrows and are explained in the text.

2. Pre-requisites

2.1. ORENTATION AND LOCATION OF SYMMETRY ELEMENTS

Success of the method depends on use of the correct symmetry operators. It is only occasionally that the initial phases will be of such quality that this is best done by direct comparison of electron density values. Usually, the transformations are determined from the diffraction data using variants of the rotation and translation functions that are used for molecular replacement [1]. By searching in Patterson space, it is possible to split a computationally intractable 6-dimensional search into a 3-D orientation search [4-6] followed by a translational search [7]. This factoring results from an assumption that is only approximately true – that peaks close to the origin (short interatomic vectors) are only dependent on the molecular orientation, and that distant peaks depend only on the relative positions of molecules. The inexactness of this and other approximations, combined with the inherent noise of Patterson functions can lead to considerable ambiguity in the interpretation of rotation, and especially translation functions. Optimal strategies of overcoming these problems are covered in other sections of this volume. Here, it is noted only that the detection of spurious solutions is more difficult than with molecular replacement, because it may be many months before it is possible to build an atomic model – special care is worth the effort!

Since the last review [3], improvements in rotation and translation functions have made the process at least slightly easier. *AMoRe* is a particularly fast implementation of both rotation and translation functions [8] that facilitates systematic examination of the interdependence of rotation and translation functions. Advances targeted towards molecular replacement -- in the sensitivity and convergence radius of searches, in rigid refinement protocols and in accounting for domains that may have different orientations [9, 10] help in cases where averaging is to be used to improve molecular replacement phase determinations. In such cases, one's confidence can be increased by checking that cross-rotation functions between model and each molecule in the unknown crystal(s) are consistent with self(/cross)-rotation functions for the diffraction data set(s) that give the relative orientations of molecules in the unknown crystal(s). Within a single crystal form, NCS is often closed point group. "Locked" rotation and translation functions apply the point group as a constraint and lead to very significant gains in sensitivity and precision [11, 12].

2.2. MOLECULAR ENVELOPE

Unlike crystallographic symmetry, NCS operators are local and do not apply to the whole unit cell. It is therefore necessary to define which parts of the unit cell should be averaged with each other. With very poor initial phases, even the distinction between protein and solvent may be difficult. Distinguishing neighboring molecules that might be operated on by different local symmetry is even more challenging.

For a large number of NCS cases, in which molecules are related by a closed point group, the challenge is lesser. It is necessary only to define an envelope containing the whole assembly, because the same operations will apply to all points within the assembly. Requirements are only that an envelope include all of the protein, and is consistent with the crystallographic symmetry.

Recent progress is in two areas. One implementation of averaging, *ENVELOPE* [13], has demonstrated the feasibility of mask refinement during averaging. The values of points in the map are compared to other points to which they might be related with adjustment of the envelope. With high order NCS, biases due to prior mis-assignment can be overcome. Both *ENVELOPE* and *MAMA* [14] (for example) have iterative procedures for forcing masks to be consistent with crystallographic symmetry, smooth and unfragmented, while the latter interfaces with interactive graphics programs for manual adjustment, and has routines to help define masks, which with appropriate transformation are consistent with two crystal forms, a requirement of averaging between crystals.

2.3. STRUCTURE AMPLITUDES & STARTING PHASES

The power of averaging comes from over-sampling of the molecular transform. If the N x fraction of data observed (N = order of NCS) falls below 1, then little phase information can be obtained. The relative effects of partial data, error *etc.* have been estimated elsewhere [15]. Exact equivalence of related points in a map can only be expected with error-free complete data. The absence of reflections has all of the bad consequences of arbitrarily setting their amplitudes to zero. Partial data sets are a fact of life for virus crystallographers, and after much argument, it has become common practice to fill in missing observations with those calculated from the back-transform of the most recent averaged map. Viruses are atypical in fraction of data typically measured, and in their higher order NCS leading to good estimates for missing observations. The author is not aware of systematic tests of filling for more typical protein structure determinations.

3. **Algorithms**

3.1. NCS AS DENSITY MODIFICATION OR AS A RECIPROCAL-SPACE CONSTRAINT

It has long been thought possible that an equivalent procedure could be applied exclusively in reciprocal space [16, 17], and its feasibility has recently been demonstrated with the re-determination of bacteriophage ϕX174 phases [18]. Due to their computational complexity, reciprocal-space algorithms have not yet widely replaced their real-space counterpart. The reader is referred to Tong & Rossmann [18] for a more complete treatment, but here an (over-) simplified summary will be given, because it is helpful in understanding properties of the real-space algorithms.

The phased (vector) structure factor, $\vec{F}$ can be calculated from a sum of other structure factors. For the simple case of N equivalents related by non-crystallographic point group symmetry operators, T_n, about the origin in space group P1:

$$\vec{F}_{h1} \propto \sum_{h2} \vec{F}_{h2} \sum_{n}^{N} G\{h_1, h_2, T_n\} \qquad (1)$$

where h_1, h_2 are reflection indices, and G is a complicated integral over the molecular volume. Note that equation 1 is a formula for iteration: improved phases can be calculated providing that there are already approximations for at least some of the phases. However, it has been suggested [19] that it may be possible to extend from F_{000} by arbitrarily setting it to one of the centric possibilities[1]. G is a correlation function between the structure factors that is similar to the interference function of rotation functions [5]. Simplification to a G-function appropriate for a spherical molecular envelope was adequate for ϕX174 (but perhaps not for proteins with lower symmetry), and is conceptually instructive. Although not necessarily monotonically, the magnitude of G decreases with $\Delta h = |h_1 - h_2[T_n]|$ (Figure 2), i.e. the phase of a reflection depends mostly on the structure factors that are neighbors after symmetry transformation.

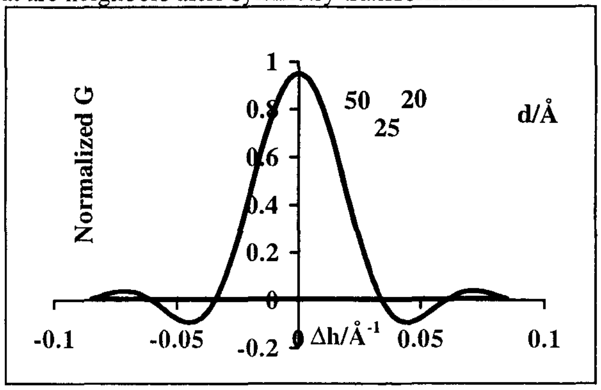

Figure 2: The normalized G correlation function for a spherical molecule of 20 Å radius is the Fourier transform of a sphere. The x-axis corresponds to the magnitude of a difference vector between two points in reciprocal space.

3.2. ALGORITHMS

3.2.1. *Averaging & Incorporation of other constraints*

With more powerful computers, averaging has ceased to be a computational challenge for all but the large virus unit cells (covered in [20]). A new generation of programs has come to the fore, including *RAVE* [21], *SQUASH* [22] and *DM* [23]. Improvements are a sign of a maturing field: user-friendliness, portability, robustness, generality, integration (combining several steps into a single program) and combination of averaging with other methods of phase improvement (*e.g.* Sayre's equation [24]) and density modification [25], such as solvent flattening [26] and histogram-matching [27]. Most recently, support has been added for multiple crystal forms and domains with different relative orientations [28, & Cowtan, K. personal communication].

In general, the NCS-equivalent of a grid point falls between other grid points, and its electron density needs to be interpolated. The precision of interpolation is increased with more finely spaced grid points. Although Shanon's theory [29] shows that no benefit results from the calculation of the *averaged* map much finer than ½ the resolution limit, Bricogne [2] showed that when linear interpolation was used, the *pre-*

---

[1] Such an extension might have equal chances of generating enantiomeric or Babinet solutions – see accompanying article [20]

averaged map needed to be sampled with a spacing finer than 1/5th resolution limit to avoid serious errors. The large unaveraged maps are now commonly avoided through the use of high-order (e.g. 11 point) interpolation through which accuracy can be achieved without fine grids. Aliasing problems are overcome with grids slightly finer than the Shannon limit, and it is now common to compute both maps (for convenience) with a grid 1/2.5 to 1/3.0 times the resolution limit.

3.2.2. Phase recombination and weighting

Phases calculated by back transformation have no figure of merit, and no width to their (unimodal) phase probability distributions. How does one calculate the map of least error? If initial phase information is worth keeping, how should the back-transformed phases be combined with it? Perhaps maximum likelihood methods will provide rigorous answers, but here let us consider methods that are currently widely used.

Most methods assume that inaccuracy of the back-transformed phases is likely to be correlated with inaccuracy of calculated structure amplitudes. Although this may be true only sometimes for individual reflections, overall, it seems reasonable. Focusing on individual relfections, aribitrary weights, w, have been used [30]:

$$w = \exp\left\{-\left|F_o - F_c\right| / \left\langle F_o \right\rangle\right\} \qquad (2)$$

Bricogne [2] adapted Sim-weighting [31, 32] to the approximation of back-transformed probability distributions:

$$P(\alpha) \propto \exp\left\{\left[2F_oF_c\cos(\alpha - \alpha_c)\right] / \left\langle F_u^2 \right\rangle\right\} \qquad (3)$$

where o, c and u subscripts refer to observed, calculated and unknown components, and where the unknown component is estimated from the average difference between F_o and F_c. Combinations of these two weighting schemes have been particularly successful [33]. Filled reflections are currently usually given the mean weight of observed reflections at that resolutions, although this is likely over-weighting these reflections. Recent programs have been enhanced through incorporation of σ_A-weighting [34] that should lead to improved weights and map coefficients, especially when partial structure is part of the initial phasing, but also when not.

Finally, a distinction has recently been made between the majority of current phase refinement methods that are designed to progress smoothly toward convergence, and search algorithms that evaluate discretely different phase sets [35]. The latter may be an exception, as it may not be beneficial to down-weight the effects of reflections with biggest disagreements.

4. Quality Assessment

A common method of monitoring progress is correlation (or residuals) of the symmetry related electron density. Monitoring in reciprocal space is also common with calculation of correlation coefficients or R-factors between observed structure amplitudes, and those back-transformed from the current map. Unfortunately, it is easy to obtain a map with near-exact NCS, even when there are errors that should be corrected. Numerical values should appear much "better" than for comparisons to atomic model, and should be "better" for low-order NCS. Use of correlation coefficients as a diagnostic indicator is covered in the accompanying article [20].

Recent advances include methods of cross-validation [36] that may provide better indicators of quality. In real-space, a sphere (and its symmetry equivalents) may be omitted from any density modification or averaging and can be used for "free" statistic evaluation [37]. Similarly in reciprocal space, a portion of the data can be omitted prior to map calculation, and this data can be used for calculation of statistics (and improved weighting coefficients) [38]. To avoid the problems introduced by omission of data, the process is repeated so that one cycle is used for statistics, another for phase calculation.

5. Applications, Further Considerations and Proof of the pudding...

Early applications in the 1960's did little to convince the crystallographic community of the general worth of NCS-phasing. With the derivation and implementation of real-space methods [2, 39], it became possible in the 1970's to determine the atomic structures of two viruses [40, 41] by improving poor initial phases. This began exponential growth of the use of NCS in phase refinement.

Although phase improvement became accepted quickly, there was much skepticism about the use of NCS to extend phases beyond the resolution of the experimental phases. These doubts were only laid to rest when NCS was used to generate maps that could be interpreted with atomic models, having started at resolutions where this would not have been possible. Haemocyanin and human rhinovirus 14 [42, 43] were among the earliest examples. What is the basis for the generation of phases where none previously existed?

Considered in reciprocal space, for a reflection, $\vec{h}$, just beyond the current phasing limit, equation 1 has half of the terms required for a precise determination of its phase. Each of the unphased reflections will have weight of zero, so $\Delta\vec{h} = 0$ will not contribute, as will the ~ ½ of the immediately neighboring reflections that happen to be at higher resolution (Figure 1). The nearly ½ of the neighbors for which phases have previously been estimated will allow a very crude approximation of $\vec{F}_h$. Addition of even a crude initial approximation for α_h facilitates its refinement, because the summation of equation 1 is more complete. Following convergence, it can also be used in the phase estimation of $\vec{h}' \geq \vec{h}$, providing a basis for phase extension.

Extension, once shown feasible, has been extended to lower and lower resolution, culminating in cases, reviewed in [20], with high order NCS where it has been possible to refine and extend from low resolution *ab initio* phases in the absence of experimental phase information. The processes of extending over a large or modest resolution range are essentially the same. One additional challenge of the *ab initio* applications is knowing that the phase determination is correct at resolutions lower than those at which molecular features should be recognizable.

The reciprocal-space understanding demonstrates the importance of several experimental factors, especially in the case of phase extension. It is important to include as many of the terms in equation 1 as possible. By increasing the resolution in steps ≤ 1 reciprocal lattice unit, nearly half of the terms can be included. Although steps of 2 or 3 reciprocal lattice units have been used occasionally for computational expediency, usually this generates poorer results, because equation 1 is missing not only all of the higher resolution terms, but some of the lower resolution terms. Clearly, anything that further detracts from the initial approximation of equation 1 is deleterious. Extension may not be able to recover from incorrect starting phases. Figure 1 shows that it is highest resolution shell that is most critical, so it is necessary to start extension at a resolution for which there is confidence in the starting phases. Incomplete or imprecise amplitudes, can be seen to affect extension through equation 1, and it may be necessary to limit the resolution to which phases are extended. Finally it is noted, that because refinement of the extended phases depends on their first approximation, it is possible for phase refinement to head towards an incorrect solution and never recover.

The usefulness of NCS is no longer questioned. With high order NCS, excellent phases can be obtained. Considerable effort has recently gone into the determination of accurate experimental phases for proteins [44] using anomalous dispersion. For the virus crystallographers lucky enough to have 60-fold NCS, although usually at lower resolution, electron density maps are routinely just as accurate [45], in spite of being calculated with amplitudes that are > 2-fold less precise.

6. Acknowledgments and References

The author is grateful to have had the opportunity to learn about averaging through research in the laboratory of one of the most preeminent of many contributors to this field, Michael Rossmann. The author is also grateful to members of his laboratory: Eric Blanc, Genfa Zhou and Qing Xie, who keep him thinking about NCS and averaging through the interesting problems their research poses, and for the financial support of the National Science Foundation and Lucille P. Markey Charitable Trust.

1 Rossmann, M. G. (1972) *The Molecular Replacement Method,* Gordon and Breach, New York

2 Bricogne, G. (1976) Methods and Programs for Direct Space Exploitation of Geometric Redundancies, *Acta Crystallographica* **A32**, 832-47

3 Rossmann, M. G. (1990) The Molecular Replacement Method, *Acta Cryst.* **A46**, 73-82

4 Rossmann, M. G. (1972) The Rotation Problem, in Rossmann, M. G. (ed.), *The Molecular Replacement Method,*, Gordon and Breach, New York, pp. 16-44

5 Rossmann, M. G. and Blow, D. M. (1962) The Detection of Sub-Units Within the Crystallographic Asymmetric Unit, *Acta Crystallographica* **15**, 24-31

6 Rossmann, M. G. and Blow, D. M. (1964) The Relative Positions of Independent Molecules Within the Same Asymmetric Unit, *Acta Crystallographica* **17**, 338-42

7 Crowther, R. A. and Blow, D. M. (1967) A Method of Positioning a Known Molecule in an Unknown Crystal Structure, *Acta Crystallographica* **23**, 544-8

8 Navaza, J. (1994) *AMoRe*: an Automated Package for Molecular Replacement, *Acta Crystallographica* **A50**, 157-163

9 Brünger, A. T. (1997) Patterson Correlation Searches and Refinement, *Methods in Enzymology* ***in press***

10 Zhang, X.-J. and Matthews, B. W. (1994) Enhancement of the Method of Molecular Replacement by Incorporation of Known Structural Information, *Acta Crystallographica* **D50**, 675-686

11 Tong, L. and Rossmann, M. G. (1990) The Locked Rotation Function, *Acta Crystallographica* **A46**, 783-792

12 Tong, L. (1996) The Locked Translation Function and Other Applications of a Patterson Correlation Function, *Acta Crystallographica* **A52**, 476-79

13 Rossmann, M. G., McKenna, R., Tong, L., Xia, D., Dai, J.-B., Wu, H., Choi, H.-K. and Lynch, R. E. (1992) Molecular Replacement Real-Space Averaging, *J. Appl. Cryst.* **25**, 166-80

14 Kleywegt, G. J. and Jones, T. A. (1993) Masks Made Easy, in (ed.), *ESF/CCP4 Newsletter,*vol. 28, 56-59

15 Arnold, E. and Rossmann, M. G. (1986) Effect of errors, redundancy, and solvent content in the molecular replacement procedure for the structure determination of biological macromolecules , *Proc. Natl. Acad. Sci. USA* **83** , 5489-93

16 Main, P. and Rossmann, M. G. (1966) Relationships among Structure Factors due to Identical Molecules in Different Crystallographic Environments, *Acta Crystallographica* **21**, 67-72

17 Crowther, R. A. (1967) A Linear Analysis of the Non-Crystallographic Symmetry Problem, *Acta Crystallographica* **22**, 758-764

18 Tong, L. and Rossmann, M. G. (1995) Reciprocal-Space Molecular-Replacement Averaging, *Acta Crystallographica* **D51**, 347-353

19 Tong, L. and Rossmann, M. G. (1995) Reciprocal-space molecular-replacement averaging, *Acta Crystallographica* **D51**, 347-53

20 Chapman, M. S., Blanc, E., Johnson, J. E., McKenna, R., Munshi, S., Rossmann, M. G. and Tsao, J. (1997) Use of non-crystallographic symmetry for *ab initio* phasing of virus structures, in Fortier, S. (ed.), *NATO Advanced Study Institute on Direct Methods for Solving Macromolecular Structures,* Kluwer Acad. Publ., Dordrecht, The Netherlands

21 Jones, T. A. (1992) a, yaap, asap, @£? A set of averaging programs, in Dodson, E., Gover, S. and Wolf, W. (ed.), *Proceedings of the CCP4 Study Weekend,*, Science and Engineering Research Council, 91 - 105

22 Zhang, Y. J. (1993) SQUASH - combining constraints for macromolecular phase refinement and extension, *Acta Crystallographica* **D49**, 213-222

23 Cowtan, K. D. and Main, P. (1993) Improvement of macromolecular electron-density maps by the simultaneous application of real and reciprocal space constraints, *Acta Crystallographica* **D49**, 148-57

24 Sayre, D. (1952) The Squaring Method: a New Method for Phase Determination, *Acta Crystallographica* **5**, 60-65

25 Cannillo, E., Oberti, R. and Ungaretti, L. (1983) Phase Extension and Refinement by Density Modification in Protein Crystallogaphy , *Acta Cryst.* **A39** , 64-74

26 Wang, B. C. (1985) Resolution of the Phase Ambiguity in Macromolecular Crystallography , *Methods in Enzymology* **115** , 90-112

27 Lunin, V. Y. (1994) Electron-density histograms and the phase problem, *Acta Crystallographica* **D49**, 90-99

28 Schuller, D. J. (1996) MAGICSQUASH: More Versatile Non-crystallographic Averaging with Mulitple Constraints, *Acta Crystallographica* **D52**, 425-34

29 Shannon, C. E. (1949) Communication in the presence of noise, *Proc. Inst. Radio Engrs. New York* **37**, 10-21

30 Rayment, I., Baker, T. S., Caspar, D. L. D. and Murakami, W. T. (1982) Polyoma virus capsid structure at 22.5 Å resoution , *Nature* **295** , 110-15

31 Sim, G. A. (1959) The distribution of phase angles for structures containing heavy atoms., *Acta Crystallographica* **12**, 813-15

32 Sim, G. A. (1960) A note on the heavy atom method , *Acta Cryst.* **13** , 511-12

33 Arnold, E. and Rossmann, M. G. (1988) The Use of Molecular-Replacement Phases for the Refinement of the Human Rhinovirus 14 Structure, *Acta Crystallographica* **A44**, 270-282

34 Read, R. J. (1985) Improved Fourier Coefficients for Maps Using Phases from Partial Structures with Errors, *Acta Crystallographica* **A42**, 140-9

35 Miller, S. T., Hogle, J. M. and Filman, D. J. (1996) A Genetic Algorithm for the *Ab Initio* Phasing of Icosahedral Viruses, *Acta Crystallographica* **D52**, 235-251

36 Brünger, A. T. (1992) Free R value: a novel statistical quantity for assessing the accuracy of crystal structures , *Nature* **355** , 472-5

37 Abrahams, J. P. and Leslie, A. G. W. (1996) Methods Used in the Structure Determination of Bovine Mitochondrial F_1 ATPase, *Acta Crystallographica* **D52**, 30-42

38 Cowtan, K. D. and Main, P. (1996) Phase combination and cross validation in iterated density-modification calculations, *Acta Crystallographica* **D52**, 43-48

39 Bricogne, G. (1974) Geometric sources of redundancy in intensity data and their use for phase determination, *Acta Crystallographica* **A30**, 395-405

40 Bloomer, A. C., Champness, J. N., Bricogne, G., Staden, R. and Klug, A. (1978) Protein disk of tobacco mosaic virus at 2.8 Ångstrom resolution showing the interactions within and between subunits, *Nature* **276**, 362-8

41 Harrison, S. C., Olson, A., Schutt, C. E., Winkler, F. K. and Bricogne, G. (1978) Tomato bushy stunt virus at 2.9 Å resolution , *Nature* **276** , 368-373

42 Gaykema, W. P. J., Hol, W. G., Vereijken, J. M., Soeter, N. M., Bak, H. J. and Beintema, J. J. (1984) 3.2 Å structure of the copper-containing oxygen-carrying protein Panulirus interruptus haemocyanin, *Nature* **209**, 23-29

43 Rossmann, M. G., Arnold, E., Erickson, J. W., Frankenberger, E. A., Griffith, J. P., Hecht, H., Johnson, J. E., Kamer, G., Luo, M., Mosser, A., Rueckert, R., Sherry, B. and Vriend, G. (1985) Structure of human common cold virus and functional relationship to other picornaviruses. , *Nature* **317** , 145-153

44 Burling, F. T., Weis, W. I., Flaherty, K. M. and Brünger, A. (1996) Direct Observation of Protein Solvation and Discrete Disorder Using Experimental Crystallographic Phases, *Science* **271**, 72-77

45 Zhou, G., Wang, J., Blanc, E. and Chapman, M. S. (1997) Determination of the Relative Precision of Atoms in a Macromolecular Structure, *Acta Crystallographica* ***submitted***