Acta Cryst. (1978). A**34**, 1014–1019

Difference Fourier Refinement of Metaquohemerythrin

BY RONALD E. STENKAMP, LARRY C. SIEKER AND LYLE H. JENSEN

Departments of Biological Structure and Biochemistry, University of Washington, Seattle, WA 98195, *USA*

(*Received* 15 *March* 1978; *accepted* 19 *June* 1978)

A model for metaquohemerythrin from *Themiste dyscritum* has been refined in the crystallographic sense of the term by difference Fourier methods at 2·8 and 2·5 Å resolution. Fourteen cycles of refinement reduced R from an initial value of 0·385 for the 9461 reflections from 10 to 2·8 Å to 0·253 for 16 363 reflections from 10 to 2·5 Å resolution. On the basis of peaks in difference maps, 49 water molecules have been added to the model for a total of 3833 atoms in the asymmetric unit.

Introduction

Hemerythrin is an oxygen transport protein found in several marine invertebrates. The molecule consists of eight chemically identical subunits, each of weight 13 500 daltons. The subunit binds two Fe atoms which are required for biological activity, but no porphyrin groups are found in hemerythrin, the Fe atoms being bound by amino-acid side chains. A smaller molecule, myohemerythrin, related to the subunit of hemerythrin, is found in the muscle tissue of organisms containing the octamer, analogous to the hemoglobin–myoglobin system found in most other organisms. Extensive discussions of the biochemistry of these molecules can be found in review articles by Klotz (1971), Llinas (1973), and Klotz, Klippenstein & Hendrickson (1976).

Crystalline samples of hemerythrin have been available since 1933 (Florkin, 1933), but low-resolution structure solutions for myohemerythrin and hemerythrin have been determined only in the past few years (Hendrickson & Ward, 1975; Ward, Hendrickson & Klippenstein, 1975; Stenkamp, Sieker, Jensen & Loehr, 1976). Published interpretations of the electron density maps (Hendrickson, Klippenstein & Ward, 1975; Stenkamp, Sieker & Jensen, 1976; Stenkamp, Sieker, Jensen & McQueen, 1978) do not agree in the identities and the arrangement of the Fe ligands. We report here the crystallographic aspects of automated difference Fourier refinement of metaquohemerythrin from *Themiste dyscritum*. A detailed description of the present model will be published elsewhere.

Data, model, and methods

Crystals grow readily from low-ionic-strength phosphate solutions at pH 7 or from 20% 2-methyl-2,4-pentanediol (Loehr, Meyerhoff, Sieker & Jensen, 1975). The space group is $P4$, with $a = b = 86{\cdot}6$, $c = 80{\cdot}8$ Å, $Z = 2$.

Data were collected from two crystals on a computer-controlled four-circle diffractometer by an ω–2θ step-scan technique (Hanson, Watenpaugh, Sieker & Jensen, 1973). The 9461 reflections to 2·8 Å resolution, for which isomorphous-replacement phases were available (Stenkamp *et al.*, 1978), were used in the initial stages of the refinement. Subsequent cycles made use of a 2·5 Å resolution data set obtained by reprocessing the original data taking into account the 2θ dependence of the radiation decay (Sheriff & Watenpaugh, 1977). Friedel pairs were averaged to yield 9805 reflections from the first crystal and 12 211 from the second. These two data sets were merged and edited to provide 16 708 reflections with $I > 2\sigma(I)$.

The initial model for this study was derived from the averaged subunit described previously (Stenkamp *et al.*, 1978) by applying the noncrystallographic symmetry in the cell to generate the four monomers in the asymmetric unit. Model 1 was composed of 8 Fe and 3772 nonhydrogen atoms.

Ten Eyck's (1973) fast Fourier routines were used to calculate the difference Fourier (ΔF) maps for the automated difference refinement (Freer, Alden, Carter & Kraut, 1975; Moews & Kretsinger, 1975; Chambers & Stroud, 1977). Difference density gradients at the atomic sites were determined by an eight-point linear interpolation of surrounding density, and full atomic shifts calculated by using curvatures of 1·0, 1·25, 1·5, 3·0 and 6·0 e Å^{-5} for C, N, O, S and Fe atoms respectively. The r.m.s. shifts in the coordinates ranged from 0·358 Å for the first cycle to 0·136 Å for the last, with maximum shifts of 1·713 and 0·875 Å respectively. After obtaining shifts from the ΔF map, we applied stereochemical constraints using the program of Hermans & McQueen (1974).

Periodically, difference Fourier maps were inspected visually to locate solvent molecules and regions of the protein requiring adjustment. This process was speeded considerably by machine plotting of the contoured density and a superimposed skeleton model of the

protein (Quigley, 1975; Anderson, Stenkamp & Steitz, 1978). The resulting sections were placed on a light box, and by leafing through the paper sheets, the locations of positive and negative density relative to the model structure were readily apparent. By avoiding the laborious task of transferring the map to clear plastic sheets, we could interpret a ΔF map and make any necessary manual adjustments of the model in a day and a half.

Refinement progress

An overall thermal parameter of 15 Å^2 was assumed in the first structure factor calculation. The trend in $\sum |F_c|/\sum |F_o|$ as a function of $\sin\theta/\lambda$ suggested an inordinately small value of B, so 8 Å^2 was selected as a reasonable compromise and was maintained throughout the refinement. R for the initial model was 0·385, a value somewhat less than usually reported for beginning models (Lipscomb, Reeke, Hartsuck, Quiocho & Bethge, 1970; Epp, Lattman, Schiffer, Huber & Palm, 1975; Ladner, Heidner & Perutz, 1977).

Fig. 1 is a plot of R ($= \sum||F_o| - |F_c||/\sum |F_o|$) *versus* model number. Reflections less than 0·05 Å^{-1} in $\sin\theta/\lambda$ (10 Å resolution), which are sensitive to the disordered solvent common in protein crystals, were omitted from the refinement and the reported R values. This eliminated 276 reflections from the initial selected 2·8 Å resolution data set and 345 from the larger 2·5 Å set. All R values, except one at the end of refinement, are for idealized models. The r.m.s. distances from the atoms in each model to their positions in the final model are also plotted in Fig. 1.

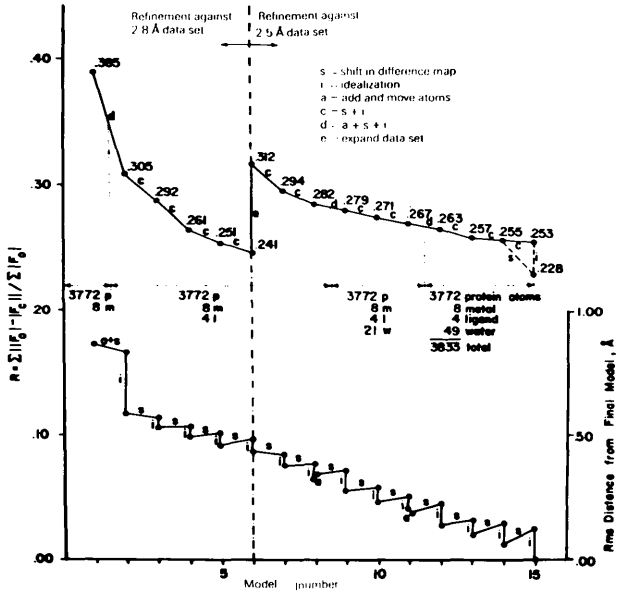

Fig. 1. R and r.m.s. distance from final model *versus* model number.

Inspection of a ΔF map calculated at an R of 0·305 indicated that additional density interpreted as water O atoms was located between each pair of Fe atoms. This was in agreement with the analysis showing these crystals to contain the metaquo form of hemerythrin (Loehr *et al.*, 1975).

Four O atoms (one per complex) were added to the first model on the basis of the difference map, and several amino-acid side chains were reoriented to fit pieces of positive density. Five cycles of automated atomic adjustment and idealization reduced R to 0·241 for the 2·8 Å resolution data set. The Fe atoms and water molecules were allowed to shift in the ΔF maps without stereochemical constraints.

At this point, the resolution was extended to 2·5 Å, R for the 16 363 reflections being 0·312. Two cycles of refinement reduced R to 0·282. A single overall scale factor had been used to this point, but it was apparent that a systematic error existed in the scaling of the data sets from the two crystals. A discontinuity in the ratio $\sum |F_c|/\sum |F_o|$ plotted in shells of $\sin\theta/\lambda$ appeared at 5 Å resolution, with the value being about 12% too high at lower resolution and 8% low at higher resolution. Data for each crystal were collected in shells of $\sin\theta/\lambda$, and 5 Å resolution corresponds to the break between the first and second shells for the first crystal. The data set was divided into two groups at $\sin\theta/\lambda = 0{\cdot}10$ Å^{-1}, and two scale factors were applied to the separate groups of reflections.

After the seventh refinement cycle, another ΔF map was calculated and inspected. 21 water molecules were located in this map and were added to the model. Only large peaks in positions appropriate for hydrogen bonding to the protein were included as water molecules. In addition, several side chains were shifted to improve the fit to the difference density.

Three additional cycles of refinement reduced R to 0·267, and 28 more water molecules were found in a ΔF map, yielding a final model containing 3833 atoms. Four more cycles of refinement reduced R to 0·253 for the idealized model. In the cycles directly preceding each of these last two ΔF map inspections, the positions of the water molecules and Fe atoms were fixed. The R value for the model before the final idealization was 0·228.

Discussion

Fig. 2 is a plot of R *versus* $\sin\theta/\lambda$ after the first and final cycles for both the 2·8 and 2·5 Å resolution refinements. The curves show the general shape found for other proteins. For the reprocessed and expanded 2·5 Å resolution data set, R increased for each shell of data to 2·8 Å, owing to the inclusion of more weak reflections. R increases disproportionately for the shell from 2·8 to 2·5 Å because the earlier refinement was against only the selected 2·8 Å data set. Note that the

2·8 Å data are fit as well after model 15 as after model 6.

No one protein-refinement technique has proven to be superior to all others. Various refinement approaches and philosophies have been developed and a much broader range of experience needs to be recorded to assess the merits of each. To aid in that assessment, we have run several tests on the final model to determine the effects of minor changes in technique on the R value. Elimination of the 10 Å resolution cut-off increases the R from 0·253 to 0·266. Hendrickson (1975) has suggested that the effect of solvent can be seen at 5 Å resolution, but in this case, the R from 5 to 2·5 Å is 0·254 compared with 0·253 for the 10 to 2·5 Å data set.

Chambers & Stroud (1977) have suggested using an empirical scale factor of the form $A_1 \exp(-A_2 x \sin\theta/\lambda) + A_3 x \sin\theta/\lambda + A_4$ to eliminate the problem of treating the solvent-sensitive low-order reflections. For hemerythrin, such a scale reduced R to 0·259 from the value of 0·266 for reflections from ∞ to 2·5 Å resolution.

Although the calculated phases for reflections with $|F_o|$ much greater than $|F_c|$ may be unreliable, the large differences carry important information. If the criterion of Chambers & Stroud (1977) is used to eliminate all reflections with $|F_o| > 4|F_c|$, the R decreases slightly to 0·251. This eliminates only 80 reflections (0·5% of the data set) for hemerythrin, compared with approximately 700 reflections (3·2%) for DIP-trypsin (Chambers & Stroud, 1977).

Protein crystals contain large amounts of solvent, the ordered part bound to the protein contributing substantially to the X-ray diffraction pattern. These solvent molecules should be included in the refined model for the best fit of the observed scattering. Considering the size of hemerythrin, the presence of an inner protein surface (*i.e.* the hole through the center of the octamer) and the relatively low thermal parameter, it is surprising that no more than about 50 water molecules were evident in ΔF maps. However, since most of the backbone amide and carbonyl groups are participating in hydrogen bonds forming the large amount of helix in the molecule, there are fewer binding sites available to the solvent. Moreover, the conservative criteria used to identify water molecules has limited the number found.

A difficulty in protein refinement involves the treatment of disordered side chains. Several side chains on the surface of the molecule were found in negative difference density with positive density nearby in reasonable positions. If the side chain was moved into the positive density, in the next ΔF map the atoms were still found in negative density with positive density appearing at the former positions, indicating that it was disordered, *i.e.* it was partially occupying both locations. The current resolution of 2·5 Å is insufficient

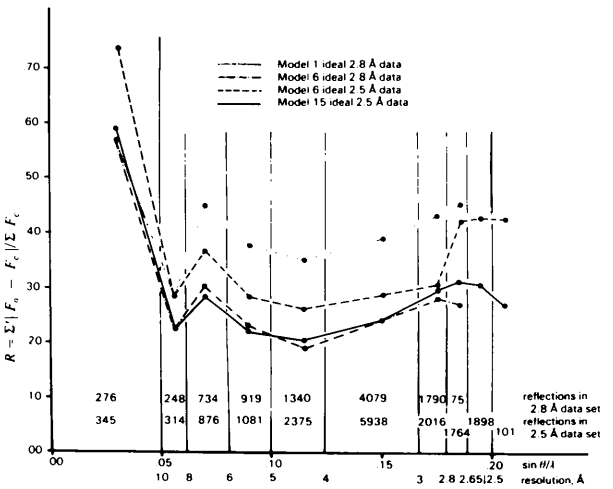

Fig. 2. R *versus* sin θ/λ for initial and final models for 2·8 and 2·5 Å data sets.

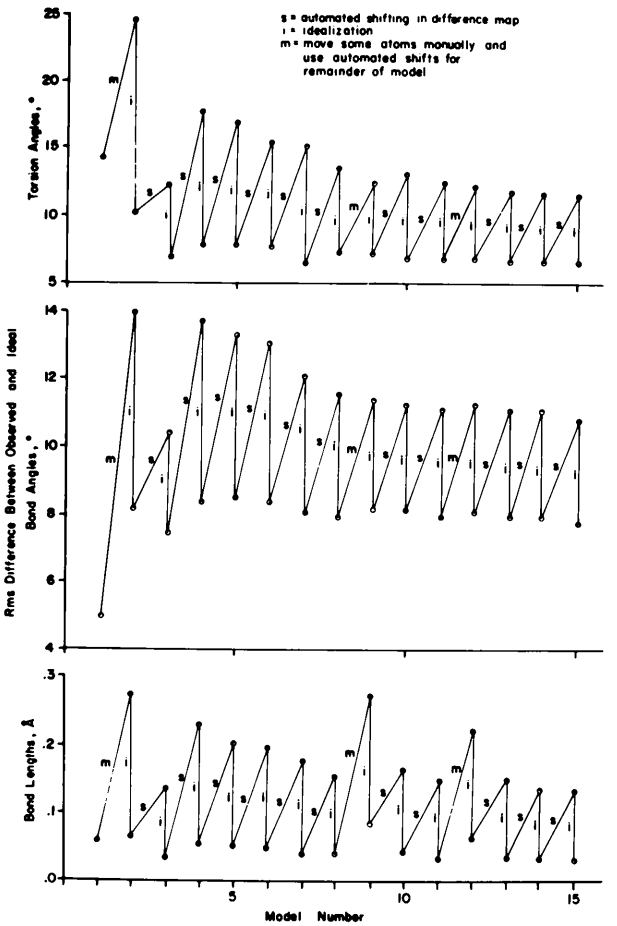

Fig. 3. R.m.s. differences between observed and ideal bond lengths, bond angles and torsion angles for the polypeptide *versus* model number.

to allow refinement of individual temperature factors, let alone partially occupied amino acid chains, so in these cases of disorder, the side chains were left in one of the alternative positions. While disorder is a reasonable explanation of our observations at this resolution, alternative models involving water molecules to fit the density cannot be ruled out.

The lower half of Fig. 1 shows the r.m.s. distance from successive models to the final one. The large shift towards the final model for the first cycle is not caused by correcting a grossly non-ideal model. The r.m.s. differences between ideal and observed bond lengths, bond angles and torsion angles for each model are plotted in Fig. 3, and two of the three do not indicate that the first model is any less ideal than later models. The large movement caused by the idealization contributed in part to the large decrease in R from 0·385 to 0·305 in the first cycle. Since the initial model was generated by making four copies of the average monomer (Stenkamp *et al.*, 1978), much of the refinement in the early cycles was simply an adjustment of the average subunits to fit the structures of the four crystallographically independent subunits in the asymmetric unit.

Based on the values given in Fig. 3, it is clear that the bond angles for the first model are more nearly ideal than those for later models, and the torsion angles less so. This could account for the large shift in the first refinement cycle and is undoubtedly caused by differences in the weighting schemes used in the idealization program.

The substantial increase in the r.m.s. error in the bond lengths observed upon manual adjustment of the model to fit the ΔF map is not reflected in the bond angles and torsion angles. It is possible that the side chains were placed in the difference density in the correct orientation, but at incorrect distances from neighboring atoms. It is more likely, however, that of the quantities plotted in Fig. 3, the error in the lengths would be more sensitive to large deviations.

The r.m.s. shifts obtained from the difference maps and from idealization are shown graphically in Fig. 4. Note that except for the first cycle, the atoms moved farther on the basis of the ΔF map than by idealization. The net shift in any one cycle is considerably smaller than the sum of the distances the atoms are moved by difference density and idealization. When comparing this method with constrained least squares (Konnert, 1976), it is important to compare only the efficiency of approaching the final model, *i.e.* the dashed-line path of Fig. 4. The ΔF and idealization steps require considerably less computer time than do least-squares methods, so the large oscillations generate a false sense of inefficiency.

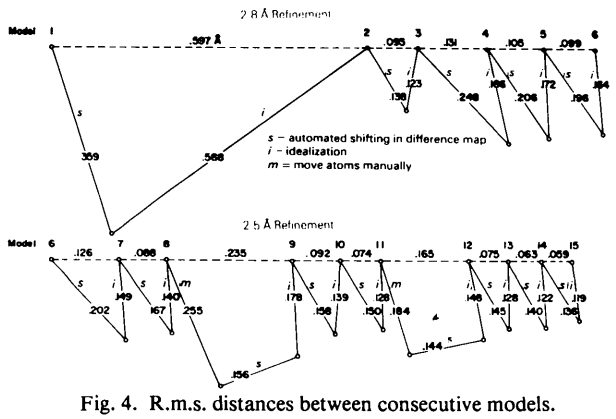

Fig. 4. R.m.s. distances between consecutive models.

In Fig. 4, it can be seen that the shifts are larger for the 2·8 Å refinement than for the 2·5 Å. This could account for the change in the r.m.s. deviations from ideality seen in Fig. 3 between models 6 and 7. The later models, after shifting in the ΔF maps, are more nearly ideal than the early ones.

The four stereoviews in Fig. 5 show the net change in the C^{α} positions from the initial to the final model. In several regions, groups of two or three residues have moved in a concerted fashion, but it remains to be determined whether the four subunits in the asymmetric unit are significantly different.

Fig. 1 shows that while R has nearly converged after 14 cycles of refinement, there is no indication that substantial shifts in the model are not yet possible. Even so, since R has essentially converged, it would be difficult to determine the relative correctness of any model obtained from further refinement with the methods described above. The r.m.s. shifts shown in Fig. 4, while steadily decreasing as the refinement progressed, also support the view that the model has not fully converged.

Although refinement has improved the fit of the calculated to the observed structure factors, we have reached a point of diminishing returns at an R of 0·253 for the data from 10 Å to 2·5 Å resolution. Since this refinement has been at relatively low resolution (other refinement studies have usually been at 2·0 Å resolution or better), we would have expected R to reach a lower value. However, the large size of the asymmetric unit and the quality of the data may contribute to the slow convergence of the refinement.

The relatively slow convergence of the refinement is related to another problem, one privately discussed by crystallographers refining proteins, *i.e.* the possibility of refining models into false minima. For small molecules, even with extensive data sets, this has been observed (Pinnock, Taylor & Lipson, 1956; Stout & Jensen, 1962; Brunton & Sears, 1969). The lack of resolution and the increased number of parameters in protein structures will aggravate this problem. Currently, there is no way of determining if the minimum found for the protein is, in fact, the true minimum, short of calculating the function being minimized for multi-

tudinous combinations of structural parameters. The relationship between the true minimum and the true structure is yet another problem.

An important result to be obtained from this refinement is the distance between the Fe atoms in the complex. The four Fe–Fe distances obtained in this study are 3·04, 3·23, 2·92 and 3·28 Å. The average value is 3·12 Å. These are all less than the value of 3·44 (5) Å reported by Hendrickson, Klippenstein & Ward (1975).

An estimated standard deviation in positon can be obtained by comparing the four subunits in the model. If the C^{α} atoms for the subunits are superposed (Ferro & Hermans, 1977) and averaged, their r.m.s. deviation from the mean positions is 0·25 Å, giving a standard deviation of 0·29 Å. This is an upper limit since the

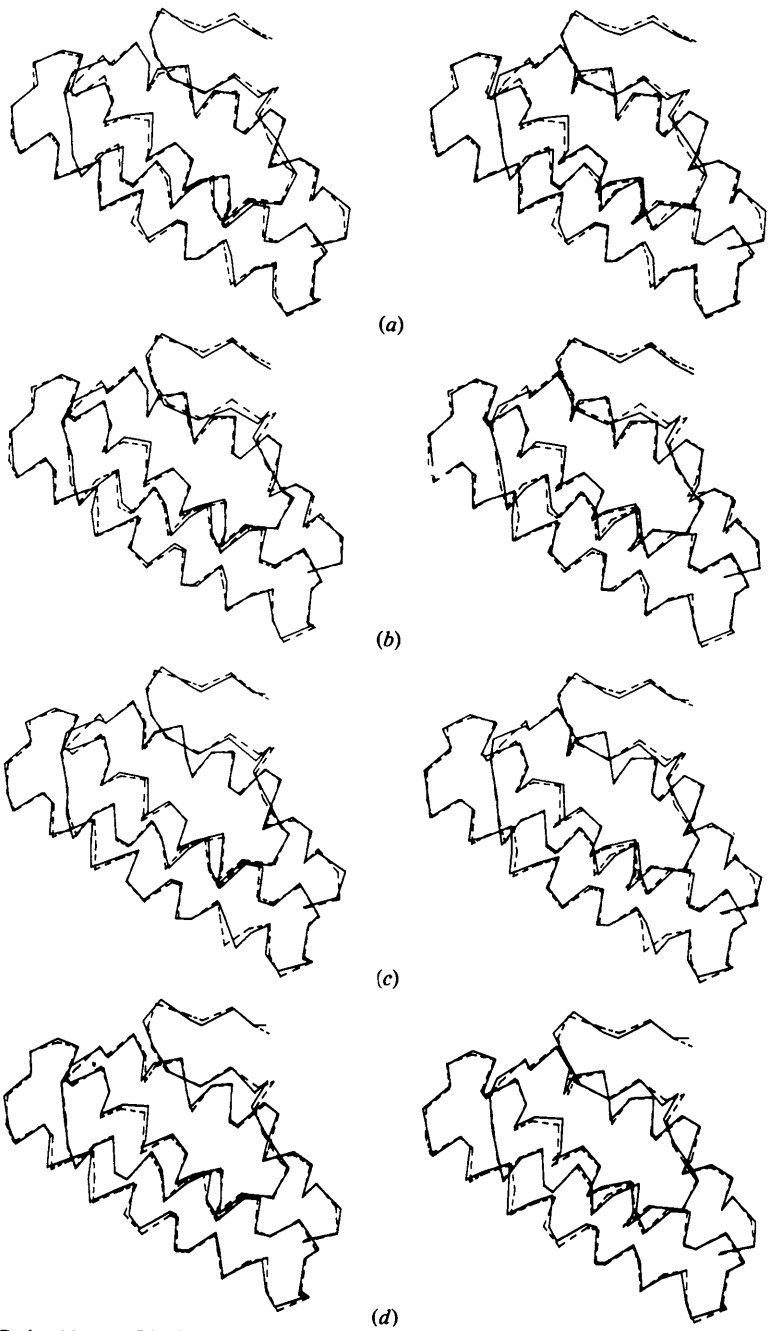

Fig. 5. Stereodrawings of C^{α} backbones. The initial model is in dashed lines, and the final in solid lines. (*a*) Subunit (I*A*), (*b*) subunit (I*B*), (*c*) subunit (II*A*), (*d*) subunit (II*B*).

assumed noncrystallographic symmetry cannot be exact. The standard deviation for the Fe atoms can be extrapolated from that of the C atoms by multiplying by the ratio of the curvatures. The value so obtained is 0·05 Å. The standard deviation in the Fe coordinates derived from the four Fe–Fe distances is 0·12 Å.

The ΔF map calculated for the final model (R = 0·253) is virtually featureless. Several more water molecules could be added to the model, but none would be located in density significantly above the noise level of the map.

The gradual decrease in R in the final cycles is characteristic of all protein refinements to date; after some rather dramatic initial decreases in R, the rate of decrease slows. There is currently no reason to think this molecule could not be refined for many more cycles, improving the model further. In the case of the Bence–Jones protein REI (Epp *et al.*, 1975), more than 30 cycles of refinement were carried out before the model was judged to be sufficiently refined. In the case of hemerythrin, the present refined model answers a number of stereochemical questions, and further refinement is not justified with the facilities and resources currently available. To answer the remaining structural questions will require a better model based on improved data extended to higher resolution.

This research was supported by USPHS Grant AM-3288 and in part by an American Cancer Society Postdoctoral Fellowship awarded to RES.

References

ANDERSON, C. M., STENKAMP, R. E. & STEITZ, T. A. (1978). *J. Mol. Biol.* **123**, 15–33.

BRUNTON, G. & SEARS, D. R. (1969). *Acta Cryst.* B**25**, 2519–2527.

CHAMBERS, J. L. & STROUD, R. M. (1977). *Acta Cryst.* B**33**, 1824–1837.

EPP, O., LATTMAN, E. E., SCHIFFER, M., HUBER, R. & PALM, W. (1975). *Biochemistry*, **14**, 4943–4952.

FERRO, D. R. & HERMANS, J. (1977). *Acta Cryst.* A**33**, 345–347.

FLORKIN, M. (1933). *Arch. Int. Physiol.* **36**, 247–328.

FREER, S. T., ALDEN, R. A., CARTER, C. W. JR & KRAUT, J. (1975). *J. Biol. Chem.* **250**, 46–54.

HANSON, J. C., WATENPAUGH, K. D., SIEKER, L. C. & JENSEN, L. H. (1972). Abstracts, Winter Meeting, Am. Crystallogr. Assoc., Univ. of Florida, Gainesville, Florida.

HENDRICKSON, W. A. (1975). Personal communication.

HENDRICKSON, W. A., KLIPPENSTEIN, G. L. & WARD, K. B. (1975). *Proc. Natl Acad. Sci. USA*, **72**, 2160–2164.

HENDRICKSON, W. A. & WARD, K. B. (1975). *Biochem. Biophys. Res. Commun.* **66**, 1349–1356.

HERMANS, J. & MCQUEEN, J. JR (1974). *Acta Cryst.* A**30**, 730–739.

KLOTZ, I. M. (1971). *Biol. Macromol.* **5**, 55–103.

KLOTZ, I. M., KLIPPENSTEIN, G. L. & HENDRICKSON, W. A. (1976). *Science*, **192**, 335–344.

KONNERT, J. H. (1976). *Acta Cryst.* A**32**, 614–617.

LADNER, R. C., HEIDNER, E. J. & PERUTZ, M. F. (1977). *J. Mol. Biol.* **114**, 385–414.

LIPSCOMB, W. W., REEKE, G. N., HARTSUCK, J. A., QUIOCHO, F. A. & BETHGE, P. H. (1970). *Philos. Trans. R. Soc. London Ser. B*, **257**, 177–214.

LLINAS, M. (1973). *Struct. Bonding (Berlin)*, **17**, 135–220.

LOEHR, J. S., MEYERHOFF, K. N., SIEKER, L. C. & JENSEN, L. H. (1975). *J. Mol. Biol.* **91**, 521–522.

MOEWS, P. C. & KRETSINGER, R. H. (1975). *J. Mol. Biol.* **95**, 201–228.

PINNOCK, P. R., TAYLOR, C. A. & LIPSON, H. (1956). *Acta Cryst.* **9**, 173–178.

QUIGLEY, G. J. (1975). Personal communication.

SHERIFF, S. & WATENPAUGH, K. D. (1977). Personal communication.

STENKAMP, R. E., SIEKER, L. C. & JENSEN, L. H. (1976). *Proc. Natl Acad. Sci. USA*, **73**, 349–351.

STENKAMP, R. E., SIEKER, L. C., JENSEN, L. H. & LOEHR, J. S. (1976). *J. Mol. Biol.* **100**, 23–34.

STENKAMP, R. E., SIEKER, L. C., JENSEN, L. H. & MCQUEEN, J. E. JR (1978). *Biochemistry*. **17**, 2499–2504.

STOUT, G. H. & JENSEN, L. H. (1962). *Acta Cryst.* **15**, 1060–1061.

TEN EYCK, L. F. (1973). *Acta Cryst.* A**29**, 183–191.

WARD, K. B., HENDRICKSON, W. A. & KLIPPENSTEIN, G. L. (1975). *Nature (London)*, **257**, 818–821.