**Dale E. Tronrud**
**Brian W. Matthews**
*Howard Hughes Medical Institute*
*University of Oregon*
*Eugene, Oregon*

# 2 Refinement of the Structure of a Water-Soluble Antenna Complex from Green Photosynthetic Bacteria by Incorporation of the Chemically Determined Amino Acid Sequence



## I. Introduction

The green photosynthetic bacterium *Prosthecochlorisaestuarii* contains a water-soluble chlorophyll-containing protein that forms part of the light-gathering antenna complex (Olson, 1978). The crystallographicstructure determination of the protein revealed it to be a trimer of three identical subunits, each containing seven bacteriochlorophyll a (Bchla) molecules (Fenna and Matthews, 1975; Matthews and Fenna, 1980). The three-dimensional crystal structure of the protein has been refined (Tronrud *et al.*, 1986), permitting the overall molecular architecture of the protein and the conformations of the Bchls to be defined fairly reliably. At the time, however, the amino acid sequence of the protein was unknown, so it was necessary to infer a tentative amino acid sequence from the crystallographic data alone. This task is not an easy one. First, the lack of an independently determined amino acid sequence makes it difficult to interpret, with confidence, the electron density maps used to define the structure of the protein. Second, even if the overall conformation of the protein is interpreted correctly, it is difficult to identify individual amino acids and to distinguish essentially isostructural pairs such as threonine and valine, glutamate and glutamine, and aspartate and asparagine.

Subsequently, however, Fenna and co-workers (Daurat-Larroque *et al.*, 1986) determined the amino acid sequence of the Bchla protein, making it

THE PHOTOSYNTHETIC REACTION CENTER, Volume I

possible to compare the X-ray sequence with the sequence determined chemically. Also it has been possible to include the reported sequence in the crystallographic refinement and, in so doing, to check the inferences that were made on the basis of the crystal structure alone. In particular, it is now possible to place the previously inferred interactions between the seven Bchls and the protein on a much more secure foundation.

## II. X-ray sequence analysis

As noted earlier, the previously published model (Tronrud *et al.,* 1986) of the Bchla protein structure was developed prior to the establishment of the chemical amino acid sequence. The X-ray sequence for the protein was based on the appearance of the electron density maps and evolved as the refinement progressed. Refinement was terminated when no apparent errors could be identified either in the positions of the atoms of the model or in the amino acid sequence. The final model (called Model 5R) contained 344 amino acids and had a crystallographic R factor of 18.9% (9 - 1.9 A resolution).

Determination of the amino acid sequence of the Bchla protein by chemical methods (Daurat-Larroque *et al.,* 1986) allows an evaluation of the prediction based on the X-ray model. Table I compares the X-ray and the chemical sequence. Overall, 74% of the amino acids were identified correctly in the electron density map. This level of success can be compared with other attempts to determine amino acid sequence from crystallographic data [e.g., 74% correct before refinement and 94% after refinement for rubredoxin (Herriott *et al.,* 1973); 53% correct for thermolysin (Colman *et al.,* 1972)]. The 95% success rate for the "high confidence" predictions, which include al-

TABLE I **Overall Accuracy of Amino Acid Sequence Prediction from X-Ray Data**

| Confidence level[a] | Total amino acids in class | Number of correct predictions | Percentage of correct predictions |
|---|---|---|---|
| *1* | *218* | *208* | *95* |
| *2* | *43* | *26* | *60* |
| *3* | *47* | *14* | *30* |
| *4* | *36* | *8* | *22* |
| Total | 344 | *256* | *74* |

[a] The "confidence level" is an empirical rating used to assess the confidence in the prediction of each amino acid. The scale is from 1 to *4; 1* is most contident. The table shows the accuracy of predictions within each category.

most two-thirds of the model, indicates that the electron density map can be used to identrfy those residues in the model for which the predicted identity is most likely to be correct. Amino acids were placed in the high confidence class by examining the quality of the image of the side chain in the map and by taking into account the location of the side chain in the structure, that is, whether it is buried or on the surface. These two criteria appear to be good predictors of the reliability of the identification of a given amino acid.

One might expect that the ability to identrfy an amino acid correctly would vary from one kind of amino acid to another. Table II summarizes the prediction success rates by amino acid type. Hydrophobic groups and groups with unique shapes are identified with high accuracy. For example, all identifications of proline, phenylalanine, leucine, tyrosine, tryptophan, arginine, and cysteine side chains proved to be correct. Amino acids identified as Asn/Asp (these were not distinguished) often were found to be leucine because the nonplanarity of the leucine side chain often was not apparent at 1.9 A resolution. Amino acids identified as Gln/Glu were correct more often because

**TABLE II Accuracy of Amino Acid Sequence Prediction by Amino Acid**

| Amino acid[a] | Number of occurrences predicted in X-ray sequence | Number of correct predictions | Percentage of correct predictions |
|---|---|---|---|
| Pro | 17 | 17 | 100 |
| Phe | 15 | 15 | 100 |
| Glu/Gln | 13 | 13 | 100 |
| Leu | 12 | 12 | 100 |
| Tyr | 9 | 9 | 100 |
| Trp | 8 | 8 | 100 |
| Arg | 8 | 8 | 100 |
| Cys | 2 | 2 | 100 |
| Ile | 23 | 22 | 96 |
| Val/Thr | 44 | 42 | 95 |
| Gly | 36 | 34 | 94 |
| Met | 5 | 4 | 80 |
| His | 9 | 7 | 78 |
| Asn/Asp | 31 | 22 | 71 |
| Ala | 48 | 18 | 38 |
| Ser | 47 | 17 | 36 |
| Lys | 17 | 6 | 35 |

[a] In the initial crystallographic refinement, no attempt was made to distinguish the essentially isostructural pairs of amino acids Glu and Gln, Val and Thr, and Asn and Asp

TABLE III **Amino Acids Predicted Correctly on the Basis of the Multiple Isomorphous Replacement Electron Density Map but Changed during Subsequent Refinement**

| Residue number in chemical sequence | Identitication in chemical sequence and MIR map | Incorrect identification after X-ray refinement |
|---|---|---|
| 10 | Thr | Ser |
| 16 | Glu | Lys |
| 30 | Lys | Ala |
| 55 | Lys | Ser |
| 78 | Lys | Ala |
| 93 | Asp | Ser |
| 109 | Ser | Ala |
| 115 | Gln | Met |
| 182 | Glu | Asx |
| 278 | Arg | Lys |
| 306 | Asp | Ser |
| 309 | Val | Ser |
| 313 | Leu | Asx |
| 360 | Phe | His |

there is no similar side chain with which they might be confused. Amino acids identified as alanine or serine were often incorrect because the actual side chain was longer, but the end was in motion and its image was not defined in the electron density map. Lysines had a poor rate of success because several arginines and tyrosines on the surface of the molecule had somewhat tenuous electron density and were predicted to be lysine.

An interesting aspect of the sequence prediction of this molecule is that 14 amino acids were identified correctly in the initial 2.8-A-resolution multiple isomorphous replacement (MIR) map (Fenna and Matthews, 1975; Matthews et al., 1979) but were changed during refinement to an incorrect amino acid. These residues are listed in Table III. A characteristic of this list is that, in almost all cases, the MIR map showed a larger side chain than was present in the final refined model. The most striking examples are the four lysine side chains that appeared clearly in the MIR map but were simply stubs of density in the $2F_o-F_c$ maps at all stages of refinement.

## III. Refinement using the chemical sequence

To improve the overall accuracy of the model, the chemical sequence information was incorporated and the refinement was continued. The amino

acids of the model were renumbered to reflect the chemical sequence; these identifications are used throughout this chapter. Each amino acid that had been predicted incorrectly was examined on an Evans and Sutherland PS330 graphics system using the program FRODO (Jones, 1982). Generally, it was necessary to place a longer side chain in place of a shorter one. The electron density was examined at a contour level of one-half the RMS value of a $2F_o$–$F_c$ map (i.e., the "0.5σ" level). In general, this map suggested a unique and chemically reasonable location in which to place the new atoms. However, without prior knowledge of the chemical sequence, there would have been insufficient confidence to place atoms in density so weak.

After a series of refinement cycles using TNT (Tronrud et *al.*, 1987), in which the chemical amino acid sequence was incorporated, the appearance of the $2F_o$–$F_c$ map improved noticeably. Many side chains that initially were placed in density at the 0.5σ contour level were now covered by density at the 1.0σ level. This result may be due to model bias but, in addition, many additional solvent molecules could be placed and the electron density in the neighborhood of the three breaks in the main-chain density became clearer.

The refinement was continued by first placing the new atoms in the model and refining this model. This process was followed by a model rebuilding and refinement loop that was repeated four times. The final model is called Model 10R and will be deposited in the Protein Data Bank. It contains 6 additional amino acids (residues 8, 58, 62, 173, 174, and 212) and 27 additional solvent atoms. Refinement statistics are given in Table N. One of the new amino acids is at the visible N terminus of the molecule whereas the others extend into the previous breaks in the chain tracing. None of these breaks have been closed, but they are significantly shorter in the present model.

As the refinement proceeded, the electron density in the neighborhood of

**TABLE IV Statistics for Refined Model 10R of the Bchla Protein**

| | |
|---|---|
| Resolution limits (Å) | 20–1.9 |
| Number of reflections | 45,335 |
| R factor (%)[a] | 17.8 |
| Weighted root-mean-square deviations from ideal values | |
| Bond length (Å) | 0.021 |
| Bond angle (°) | 3.3' |
| Planarity, trigonal (Å) | 0.023 |
| Planarity, other planes (Å) | 0.024 |
| Torsion angle (°)[b] | 19.2 |

[a] The R factor is defined as $R = \Sigma|F_o-F_c|/\Sigma F_o$, where $F_o$ and $F_c$ are the observed and calculated structure amplitudes and the sum is over all reflections. The model is used to calculate the $F_c$ includes a term to account for disordered solvent.

[b] The torsion angles were not restrained during refinement.

residue 124 (Tyr) became stronger and the main-chain conformation became clearer, but the location of the side chain remained obscure. There were strong features in the difference map ("$F_o$-$F_c$") on each side of the main chain, but this density did not seem to correspond to tyrosine or to any other amino acid. Therefore, alanine was built at this position. Also, at the adjacent amino acid (Arg 125), no density could be detected for the putative side chain, so this residue was built in the model as glycine.

Residue 117 is identified in the chemically determined sequence as glutamate but this seems to be inconsistent with the X-ray data. The previous X-ray model contained a serine at this position. When refinement was attempted with a glutamate side chain, there were apparent steric clashes and the electron density did not obviously correspond to glutamate. Therefore we suggest that the chemical sequence should be checked at this location. Some other identifications in the chemical sequence are not clearly consistent with the X-ray data (residues 68, 203, 208, 233, and 333), but the conflicts are not as clear cut as for residue 117.

Other than the changes mentioned, the model for the Bchla protein changed very little on refinement with the inclusion of the chemically determined amino acid sequence. Some structural adjustments were made at the breaks in the main chain where new amino acids were added. These additions required shifts in the atoms of the adjacent residues. For almost all of the structure, however, the shifts were less than 0.2 A, which is approximately the expected accuracy of the coordinates for a 1.9-A-resolution model.

The shifts in the Bchls were also quite small. The only substantial change was in Bchl 4, where the end of the phytyl chain was rotated by 180°. The extremities of the phytyl chains are, in any event, less well defined than the rest of the Bchla molecules; one or more of these might still be in an incorrect conformation, even in Model 10R.

## IV. Deductions from the new model

Because the initial sequence prediction based on the X-ray data alone was so successful, and the structure changed so little on subsequent refinement, almost all the conclusions based on Model 5R still stand (Tronrud *et al.*, 1986). The fold of the protein, the location of secondary structure, the relationship of the Bchls to the protein, the identity and conformation of the magnesium ligands, the conformation of the Bchla ring substituents, and the phytyl tails are all essentially the same as described previously. Specifically, the ligands to five of the Bchls are histidines (110, 145, 290, 297, and 298), whereas Bchl 2 is ligated by a solvent molecule and Bchl 5 is ligated by the backbone carbonyl of residue 242 (cf., Table 4 of Tronrud *et al.*, 1986).

The one conclusion from the previous model that is not supported by the present refinement is the segregation of the Bchla macrocycles into two classes with different types of distortion (Tronrud *et al.*, 1986). The correlation between the out-of-plane distances for some pairs of Bchla rings remains high but the correlation coefficients, in general, have decreased. One cannot conclude that the observed distortions of the macrocycles follow a distinct pattern.

Two puzzling aspects of the Bchl core might be noted. First, there is a 2.8-Å contact between atom CBC of Bchl 3 and solvent molecule SOL6. (Atom CBC is the distal carbon of the ethyl substituent on Ring C. See Fig. 6 of Tronrud *et al.*, 1986.) This distance is too short for a van der Waals contact between a water molecule and a methyl group. The electron density, however, is clear for both groups. The water molecule is buried completely and would make four hydrogen bonds in a tetrahedral fashion, if one considered the contact to atom CBC to be the fourth hydrogen bond. The presence of such a hydrogen bond would necessitate the radical assumption that atom CBC was, in fact, an oxygen or hydroxyl rather than a carbon atom. A check of the other six Bchls in the structure shows that, in five cases, the atom CBC is located in an extremely hydrophobic environment consistent with its identification as a methyl group. The two exceptions are Bchls 3 and 6. In the case of Bchl 6, atom CBC is near the $O^{\gamma}$ oxygen of residue 235 (3.9 A) and the OBB oxygen of Bchl 7 (4.0 Å) but does not make a contact as close as that observed for Bchl 3.

The precision of a 1.9-Å-resolution model does not allow one to distinguish a carbon atom from an oxygen atom, either by examination of the shape of the atom or by its bond lengths. Differences between expected bond angles and torsion angles for an oxygen relative to a carbon might just be recognizable. If we assume that the CAC–CBC atoms of Bchl 3 do not correspond to an ethyl group but to an aldehyde then the C3C–CAC–CBC angle should be 120" instead of 110". This angle for the seven Bchl rings is 104°, 109°, 120°, 114°, 114°, 117°, and 115°, respectively. The angles for Bchls 3 and 6 are the largest and are close to 120°, but the spread in values for the other rings makes the test inconclusive. The torsion angle C4C-C3C-CAC-CBC can, in principle, be used also to distinguish an ethyl and an aldehyde. In the present context, the torsion angle is a better indicator than the bond angle because it is less sensitive to the low resolution nature of the model and because torsion angles were not restrained in the refinement. For an ethyl group, the torsion angle should be 180" and for an aldehyde ±90". The observed angle for each ring is 170°, −170°, 101°, 171°, −178°, −29°, and 165". The root-mean-square discrepancies of all torsion angles from their "ideal" values is 19.2" (Table IV), which provides an estimate of the experimental error in this quantity. Again, Bchl 3 is consistent with the aldehyde identification and Bchl 6 might be. It must be stressed, however, that the torsion angle energy poten-

tial well is not very deep, and external influences might cause a significant deviation. An additional indication of difference between Bchls 3 and 6 and the others is that the phytyl tails of these two Bchls fit their density less well than the other five.

The obvious difficulty with the hypothesis that Bchl 3 or 6 does not correspond to Bchla because it contains an aldehyde side chain at this location is that no such variant of Bchla has been reported in the literature. To propose a novel type of Bchla based on such indirect and inconclusive evidence is difficult. We present the data simply to bring the possibility to the attention of the reader.

The second unusual aspect of the core is that a number of the oxygen atoms within the head groups of the Bchla moieties appear not to participate in hydrogen bonding. Although in most cases the oxygen atoms involved in the ester linkage between the Bchl head group and the phytyl tail make hydrogen bonds as expected, those of rings I and II do not. These two linkages are buried in the hydrophobic core, near each other, but there are no apparent hydrogen bond donors in thevicinity.

## V. Conclusions

One of the principal conclusions from the refinement is that the structure could be determined and refined to high resolution without knowledge of the amino acid sequence. The model so produced gave an accurate picture of the structure and the amino acid sequence deduced from it gave reliable information about the ordered core of the molecule. We attribute the success of the refinement, at least in part, to the care that was taken to include side chains in the model when it was justified by the electron density. Otherwise, errors may introduce bias into the model that is very difficult to remove.

Another conclusion is that errors are still scattered throughout the model. These caused the images of some amino acids to be obliterated. In Model 5R, several amino acids were not visible. Subsequent refinement, however, leading to Model 10R, showed that these residues could be visualized. The inability to locate these residues in earlier stages of the refinement apparently was caused by phase errors introduced by mistakes in distant portions of the model. These particular amino acids are disordered to some degree and are more vulnerable to such phase error because their electron density is weak. With reduction in phase error, they can, however, be seen. No doubt if the errors could be reduced further, additional portions of the molecule with relatively high mobility also would become visible.

Although the structure of the Bchla protein has been determined, it has not made it possible to calculate absorption and circular dichroism spectra in agreement with the spectra observed experimentally (Pearlstein and Hemen-

ger, 1978). The possible presence of a different species of Bchl might help explain this difficulty but, at present, simply adds another level of uncertainty.

## Acknowledgments

We thank Dr. John Olson for helpful advice. This work was supported in part by grants from the National Science Foundation (DMB8611084) and the Lucille P. Markey Charitable Trust.

## References

Colman, P. M., Jansonius, J. N., and Matthews, B. W. (1972). The structure of thermolysin: An electron density map at 2.3Å resolution. J. Mol. Biol. 70, 701–724.

Daurat-Larroque, S. T., Brew, K., and Fenna, R. E. (1986). The complete amino acid sequence of a bacteriochlorophyll a-protein from Prosthecochloris aestuarii. J. Biol. *Chem.* 261, 3607–3615.

Fenna, R. E., and Matthews, B. W. (1975). Chlorophyll arrangement in a bacteriochlorophyll protein from *Chlorobium limicola.* Nature (London) 258, 573–577.

Herriott, J. R., Watenpaugh, K. D., Sieker, L. C., and Jensen, L. H. (1973). Sequence of rubredoxin by X-ray diffraction. J. *Mol.* Biol. 80, 423–432.

Jones, T. A. (1982). FRODO: A graphics fitting program for macromolecules. In "Crystallographic Computing" (D. Sayre, ed.), pp. 303–317. Oxford University Press, Oxford.

Matthews, B. W., and Fema, R. E. (1980). Structure of a green bacteriochlorophyll protein. *Acc. Chem.* Res. **13,** 309–317.

Matthews, B. W., Fema, R. E., Bolognesi, M. C., Schmid, M. F., and Olson, J. M. (1979). Structure of a bacteriochlorophyll a-protein from the green photosynthetic bacterium *Prosthecochloris aestuarii. J. Mol.* Biol. 131, 259–285.

Olson, J. M. (1978). In "*The Photosynthetic Bacteria.*" (R. K. Clayton, and W. R. Sistrom, eds.), pp. 161–197. Plenum Press, New York.

Pearlstein, R. M., and Hemenger, R. P. (1978). Bacteriochlorophyll electronic transition moment directions in bacteriochlorophyll a-protein. Proc. *Natl. Acad.* Sci. *U.S.A.* **75,** 4920–4924.

Tronrud, D. E., Schmid, M. F., and Matthews, B. W. (1986). Structure and X-ray amino acid sequence of a bacteriochlorophyll a-protein from Prosthecochloris aestuarii refined at 1.9Å resolution. *J.* Mol. *Biol.* 188, 443–454.

Tronrud, D. E., Ten Eyck, L. F., and Matthews, B. W. (1987). An efficient general-purpose least-squares refinement program for macromolecular structures. *Acta Cryst.* **A43,** 489–503.