ASSIGNMENT OF SEGMENTS OF THE BACTERIORHODOPSIN SEQUENCE TO POSITIONS IN THE STRUCTURAL MAP

J. Trewhella, S. Anderson, R. Fox, E. Gogol, S. Khan, and D. Engelman
Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06511

G. Zaccai
Institut Laue-Langevin, 38042 Grenoble, France

ABSTRACT Specific amino acid sequence segments have been assigned to locations in the structural map of bacteriorhodopsin using two-dimensional neutron diffraction data and a model building analysis. Models are constructed computationally by building specific regions of the amino acid sequence as alpha helices and then positioning the helices on axes indicated by the density map of Henderson and Unwin (*Nature* [*Lond.*]. 1975, 257:28–32). Neutron diffraction data were collected from samples of stacked, oriented "native" purple membranes as well as purple membranes containing different kinds of deuterated amino acids. Models differing in the assignments of helices to specific axes and in rotations of the helices about those axes were tested against the neutron data using a weighted residual factor to rank the models. This residual factor was calculated between observed and predicted intensity differences for pairs of data sets. Using this approach, a small set of related models has been found that predicts the observed intensity changes between five independent data sets. These models are inconsistent with the proposed locations of the retinal chromophore and the carboxyl terminus and with any of the previously proposed models for bacteriorhodopsin.

INTRODUCTION

Bacteriorhodopsin is a small protein found in crystalline patches in the plasma membrane of *Halobacterium halobium.* When light is absorbed by its retinal prosthetic group, protons are translocated across the membrane out of the cytoplasm, resulting in the storage of a part of the light energy in an electrochemical gradient (for review, see reference 1). Bacteriorhodopsin's small size, crystalline organization in the purple membrane, and bioenergetic function have spurred rapid development of a number of approaches to its study. From a structural perspective, the two most significant advances have been the low resolution three-dimensional structural map of Henderson and Unwin (2) and the determination of the amino acid sequence by Khorana et al. (3) and Ovchinnikov et al. (4).

An important objective of many current studies is to arrange the amino acid sequence in a structure corresponding to the density map and consistent with known chemical modification data and stereochemical constraints. Several efforts to produce such a model have been made (4–6). These efforts led to a choice of assignment of aminò acid sequences to the helical regions presumed to span the membrane and to suggestions for the assignment of specific helices to locations in the structure. Unfortunately, the information available has not been sufficient to limit the possibilities even to a few models. Our objective is to use neutron diffraction data from specifically deuterated purple membranes to constrain the number of models.

In preliminary studies, deuterated valine and deuterated phenylalanine were biosynthetically incorporated into bacteriorhodopsin (7). Neutron diffraction data were measured, and the observed intensity changes were used, together with other information, to obtain difference Fourier maps. Using a simple model, an interpretation was made that the polar groups on the helices were predominantly toward the molecular interior, while the nonpolar groups were directed toward the surrounding lipids (7).

The described analysis was through the interpretation of difference Fourier maps and involved three important qualifications: nonequivalent reflections were overlapped in the diffraction pattern and were separated according to the ratio observed in electron diffraction patterns, and phases based on the electron microscope image were used; the difference Fourier maps used are intrinsically noisy since the structure is acentric.

Dr. Khan's current address is the Department of Biology, California Institute of Technology, Pasadena, CA. Dr. Gogol's current address is the Department of Structural Biology, Stanford University, Palo Alto, CA.

An alternative analysis using a model building approach is presented here. Regions of the amino acid sequence are assigned to helical segments, and their atoms arranged in three dimensions according to standard alpha helical parameters. Helix axis positions are determined from the electron-diffraction density map. Different models are constructed by assigning specific helices to axial positions and varying the rotational orientations of each helix about its axis. Neutron scattering intensities are then calculated for each model. By selectively using deuterium scattering factors in the calculation, intensity changes are predicted for the deuteration of each kind of amino acid. The degree to which a given model predicts the observed intensity changes is assessed by calculating a weighted residual factor.

Additional neutron data have been measured from purple membranes containing deuterated valine, deuterated phenylalanine, deuterated leucine, and deuterated isoleucine, as well as membranes with no deuterium-labeled amino acids. Using this data, and the model building analysis, we have found assignment models that predict the observed intensity changes. All of these models place helices labeled F and G in the amino acid sequence in density positions 3 and 4 in the low-resolution structural map. These assignments are inconsistent with any of the previously proposed sequence-to-density assignments for bacteriorhodopsin (4–6) as well as the proposed location of the retinal chromophore by King et al. (8) and the location of the carboxyl terminus suggested by Wallace and Henderson (9).

MATERIALS AND METHODS

Sample Preparation

Halobacterium halobium, strain S9, was grown on a defined medium as described previously (7). Cultures were grown to stationary phase at 37°C with controlled illumination and aeration. Purple membranes were isolated from washed cells by the method of Oesterhelt and Stoeckenius (10).

Deuterated isoleucine, leucine, and phenylalanine were isolated from deuterated protein according to the method of LeMaster and Richards (11). Deuterated protein was obtained from *Escherichia coli* grown in D_2O. Deuteration levels in the isolated amino acids were 85–95% as measured by ^{1}H NMR. Deuterated valine was obtained commercially.

Levels of incorporation of labeled amino acids were measured by three different techniques. Incorporation of tritiated valine in test growths of purple membranes was measured using a combined amino acid and radioactive analysis (7). ^{1}H NMR difference spectroscopy was used to measure the level of incorporation of deuterated phenylalanine in the sample used for measurement. Bacteriorhodopsin was extracted from washed, lyophilized purple membranes using a repeated acetone/ammonia extraction (12) and then solubilized in deuterated sodium dodecyl sulphate (SDS). ^{1}H NMR spectra were recorded from samples that were ~0.5 mM in protein (1% SDS) using a 500 MHz spectrometer (Bruker Instruments, Inc., Billerica, MI). Protein concentrations were determined using a standard Lowry assay, and the incorporation level was determined from a difference spectrum. Combined gas chromatography and mass spectroscopy was used to measure levels of incorporation of deuterated leucine and isoleucine. Bacteriorhodopsin was extracted from purple membranes as described for the ^{1}H NMR analysis. The extracted protein was hydrolyzed (6 N HCl, 110° for 22 h) and the resulting amino acid mixture was derivatized using a trimethylsilyl (TMS) derivatization (13). This mixture of TMS derivatives was run through a GC/mass spectroscopy analysis as described by Gehrke et al. (13). Levels of deuteration were determined from the mass distributions measured for the individual TMS derivatives that were separated on the GC column.

Multilayered specimens were prepared by drying suspensions of purple membranes (20–30 mg/ml) on acid-washed glass or quartz slides. The drying was carried out over a period of 12 h, with ~20 mg of membranes deposited over an area of 4 cm^2. Approximately 10 of these slides were stacked together to constitute one sample for neutron diffraction. Omega scans in the neutron diffractometer showed that such preparations have a mosaic spread of ~10° full width at half maximum.

Data Collection

Neutron measurements were carried out according to the method of Zaccai and Gilmore (14), using the D-16 diffractometer at the Institut Laue-Langevin in Grenoble. This instrument consists of a set of collimating Soller slits that direct neutrons of 4.60 ± 0.05 Å wavelength through the sample initially mounted with the planes of the membranes perpendicular to the incident beam. A second set of Soller slits set at an angle of 2θ to the direction of the collimating slits allows neutrons scattered in a narrow angular range (0.1°) to strike the ^{3}He detector. Data were collected in $\omega/2\theta$ scans. The sample was maintained at 25°C and 76% relative humidity (H_2O) (15).

Neutron intensity data were collected for purple membranes containing no deuterated amino acids ("native"), and for membranes containing deuterated valine (D Val), deuterated phenylalanine (D Phe), deuterated leucine (D Leu), and deuterated isoleucine (D Ile). Native, D Phe, and D Val data were each collected twice from independently prepared samples. A Lorentz factor of $(h^2 + k^2 + hk)^{1/2}$ was applied to the intensities. The difference in absorption between the two extremes of a scan was negligible, and no correction was applied. To correct for systematic differences between independent data sets measured for the same derivative, an exponential scale factor of the form $A\exp(-4B\sin^2\theta/\lambda^2)$ was applied, where A and B were the intercept and slope of the error-weighted least-square lines of best fit, calculated from plots of $\log_e (I_o^a/I_o^b)$ vs. $4\sin^2\theta/\lambda^2$. I_o^a and I_o^b are the observed intensities for each data set. B values calculated in this way were between 5 and 35 Å^2.

Formulation and Evaluation of Models

The first problem in constructing a model is to decide which parts of the amino acid sequence are present in the seven helices seen in the electron diffraction density map. Our assignment is based on a free energy calculation reported elsewhere (16), in which the partitioning of successive segments of the sequence between aqueous and nonaqueous environments is considered. From this calculation regions of the polypeptide chain presumed to span the hydrophobic region of the bilayer are assigned (Fig. 1). The hydrophobic region of the purple membrane is estimated to be 30 Å across (16), and 21 amino acids in an ideal alpha helix would just span this distance. For model building purposes each polypeptide segment, with the exception of helices D and F, was modeled as a 25–amino-acid standard alpha helix (21 amino acids to span the hydrophobic regions plus 2 amino acids at either end). Helices D and F were modeled as 26 and 27 amino acid alpha helices, respectively, to include the three valine residues located at the ends of these segments. Amino acid side chains were placed in the idealized conformations described by Diamond (17).

To complete a model, the individual helices were positioned along straight axes that best fit the rods of density observed in the three-dimensional map (2). At 7 Å resolution the rods of density are sufficiently approximated as straight for model building purposes. Different models can be constructed by assigning particular helical sections in the sequence to specific density positions (assignment models), and varying the rotational orientation of the helices about their axes (rotation models). An assignment model consists of placing the helices A–G (Fig. 1) in the

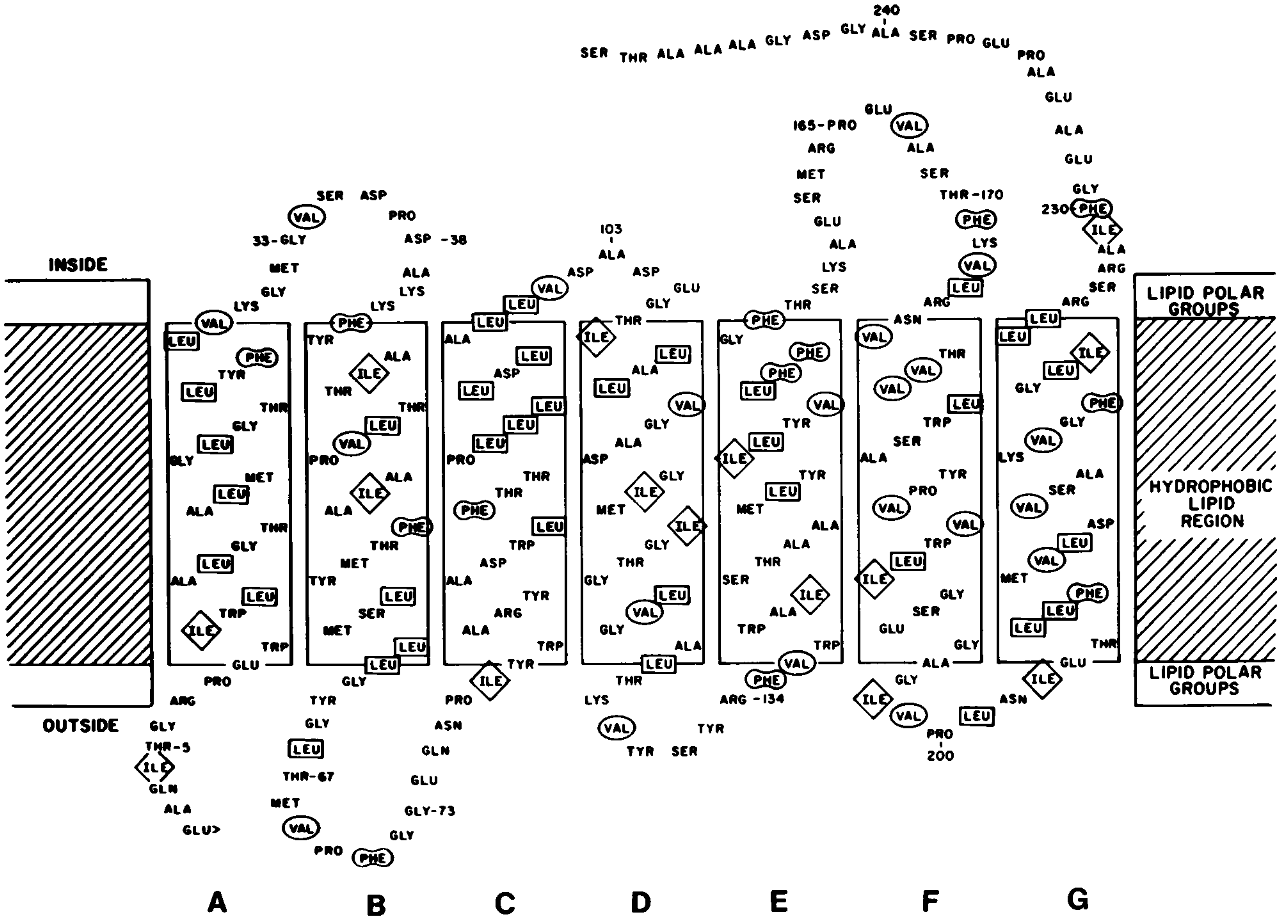

FIGURE 1 Proposed arrangement of the polypeptide chain of bacteriorhodopsin showing the helical regions presumed to span the bilayer. The helices are designated by the letters *A–G* beginning at the amino terminus. Sites of deuteration for each of the derivatives are indicated by marking the locations of valine, isoleucine, leucine, and phenylalanine.

density locations 1–7 (Fig. 2) and is identified by giving the sequence of lettered helices in the order 1–7. As there are seven helices, there are seven factorial (or 5,040) possible assignment models. Allowing n-fold rotations for the helices, this gives a total of $7! \times n^7$ models in the starting set. For $n = 6$ there are 1.41×10^9 possible models.

The vectorial contribution to the total equatorial structure factor of each helical segment (s) when positioned in each of the seven helical regions (p) and oriented in each of the equally spaced rotations about the helix axis (r) is

$$f(h, k)_{s,p,r} = a(h, k)_{s,p,r} + ib(h, k)_{s,p,r}$$

where

$$a(h, k)_{s,p,r} = \sum_i b_i \cos 2\pi(hx_i + ky_i)$$

$$b(h, k)_{s,p,r} = \sum_i b_i \sin 2\pi(hx_i + ky_i)$$

and the b_i are the neutron scattering lengths for each atom whose positional coordinates are (x_i, y_i) when projected onto a plane parallel to the membrane sheet.

Rapid computation of the calculated intensities for given assignment models was achieved by initially tabulating all values of $a(h,k)_{s,p,r}$ and $b(h,k)_{s,p,r}$. The real and imaginary components of the total structure factor $F(h,k)_c$ were computed as:

$$A(h, k)_c = \sum_{i=1}^{7} a(h, k)_{s,p,r}$$

and

$$B(h, k)_c = \sum_{i=1}^{7} b(h, k)_{s,p,r}$$

where p and r are the individual helix positions and rotation assignments for the specific model. Model intensities were calculated as $I(h,k)_c = A(h,k)_c^2 + B(h,k)_c^2$ and summed with those calculated for overlapping reflections to yield a calculated estimate of each observed intensity, I_{ci}. Calculated intensities were determined in this way for the native structure and for each derivative.

As a simple test of the basic structure, a conventional residual was calculated between observed and calculated electron microscopic structure factor amplitudes $[\Sigma_i(F_{ci} - F_{oi})/\Sigma_i(F_{oi})]$. Values obtained for different assignment models were in the range 29–34%.

The observed neutron intensity data were put on an absolute scale for comparison with the calculated data by applying a linear scale factor, $k = \Sigma_i I_{ci}/\Sigma_i I_{oi}$ to the observed data, where I_{oi} and I_{ci} are the observed and calculated intensities. This scale factor was calculated for each model. The 1,0; 1,1; 2,0 and 2,1/1,2 intensity peaks were not included in the scaling, nor were they used in subsequent calculations. This eliminated the need to correct for systematic differences observed between calculated

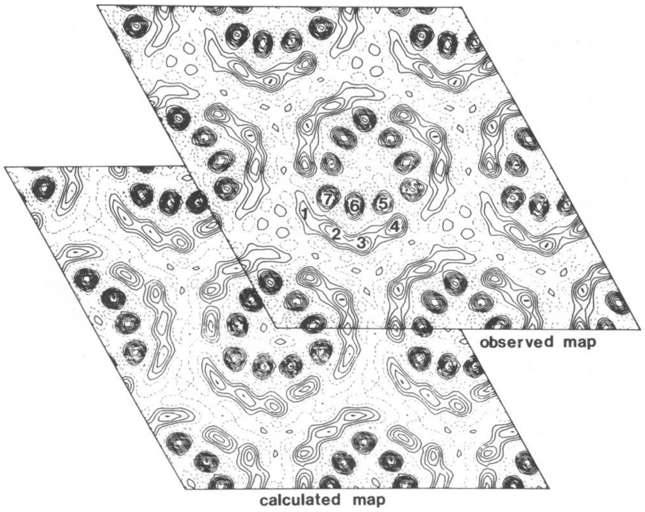

FIGURE 2 Comparison of the projections of the observed and calculated electron density map. The observed map is based on the 7-Å electron microscope data. The calculated map is based on structure factors and phases calculated for model AGFEDCB. No significant changes were observed in the calculated projections on variation of the sequence-to-density or rotation assignments. Both maps were calculated using the same number of reflections and the data sets were scaled to make the sums of the squared structure factors equal. The convention used for numbering the seven regions of presumed helical density is shown on the observed map. Assignment models are designated by giving the helix letters in the order of the numbered density positions shown.

intensities as a function of scattering angle. Plots of $\log_e(I_c/I_o)$ vs. $4 \sin {}^2\theta/\lambda^2$ for different models showed little or no systematic falloff of the observed data beyond the 1,2/2,1 intensity peak.

Models were assessed with the aid of a weighted residual factor defined as

$$R = \left[\frac{\sum_i \frac{1}{w_i} (\Delta I_{ci} - \Delta I_{o_i})^2}{\sum_i \frac{1}{w_i} (\Delta I_{o_i})^2} \right]^{1/2}$$

where ΔI_{oi} and ΔI_{ci} are the observed and calculated intensity differences between two different data sets and $w_i = \sigma_i^2 + K$. The standard deviation, σ_i, associated with each ΔI_{oi} is calculated from counting statistics while K is a constant that approximates the systematic errors in the data. K was estimated by calculating the mean value of the deviations between the observed and calculated intensity differences over a large number of models for each data pair. It was typically of the same order as the mean σ_i and was not varied from model to model. In test calculations where the w_i were set equal to 1 or σ_i^2, it was found that the R values increased, but that the rank ordering of models was preserved.

It should be noted that the R test for the relative merit of different models is based on a comparison of observed and calculated differences for intensity peaks that arise from overlapped nonequivalent reflections. A number of combinations of intensities for the overlapped reflections could give rise to the same total intensity change. It is possible, therefore, to obtain a small R factor for incorrect models, although the correct model should also have a small R. The availability of several derivatives increases the probability that a correct model will be the only solution from a number of independent experiments.

All the programs for the model building calculations were written in FORTRAN to run a PDP11/70 computer (Digital Equipment Corp., Maynard, MA).

RESULTS AND DISCUSSION

Table I shows the neutron intensity data collected from native purple membranes and purple membranes with labeled amino acids. Changes in the relative intensities are evident between the different data sets. There are 22 measurable intensity peaks in a data set, corresponding to 35 independent reflections to a resolution of ~7 Å. Because the sample consists of a stack of rotationally disordered sheets, the ***h*,*k*** and ***k*,*h*** reflections are overlapped. These reflections are not, in general, equivalent, since the symmetry of the lattice is p3.

Table II gives the incorporation levels measured for labeled amino acids supplied in the growth medium. Incorporation of the label into other amino acids was never more than 2%.

The Significance of R

To gain some insight into the significance of the calculated R values, a number of test calculations were done. The smallest expected R value, R_o, based on the errors in the data alone, was calculated by putting $(\Delta I_{oi} - \Delta I_{ci})$ equal to stochastic errors, E_i, generated from the most probable errors in the data. $E_i = \sigma_i x_i$, where the x_i are random numbers such that $-1.348 < x_i < 1.348$, thus the average magnitude of E_i is equal to the expected probable error, 0.674σ. R_o values of ~20% were found for each of the observed data pairs.

R factor calculations were also done for selected model using error modified observed data, $\Delta I_{oi} + E_i$. Over a number of trials, R values for "good" models ($R_a \sim$ 20–30%) were found to be generally larger (by up to 10%), but the rank ordering of assignment models was unchanged and the favored rotation models were very similar to those obtained with the actual observed data, differing only by rotations of one or two helices by 60°.

The observed increase in the R values when errors are added to the data is expected. If R_a is the R factor for a given model when tested against a given pair of data sets, and $R_{a'}$ is the R factor when the observed intensity differences have been modified by a set of stochastic probable errors, it can be shown that $R_{a'}^2 \sim R_a^2 + R_o^2$. For good models, with $R_a \sim R_o \sim 20\%$, $R_{a'} \sim 28\%$. Thus, models with R values between 20 and 28% would not be distinguished by the R test, given the errors in the data. For larger values of R_a, the range of R values over which models cannot be distinguished decreases, e.g., if $R_a \sim$ 40%, $R_{a'} \sim 45\%$.

For the model building analysis, we have taken a difference in R value of 10% to be significant in distinguishing between models. From our understanding of the behavior of the R values with respect to the errors in the data, this figure would seem conservative.

The above arguments do not take into account errors arising from deficiencies in the starting model, such as the absence of the linking regions, the degree of nonideality in

TABLE I
LORENTZ CORRECTED INTENSITY DATA WITH STANDARD DEVIATIONS

h, *k*	Native	D Val	D Phe	D Ile	D Leu
1,0	3,736 (556)	3,534 (529)	4,551 (670)	9,554 (414)	10,525 (357)
1,1	36,695 (886)	38,440 (840)	47,200 (572)	36,145 (1,483)	68,485 (2,085)
2,0	14,512 (688)	12,651 (575)	17,480 (865)	18,236 (1,262)	17,743 (1,156)
1,2;2,1	9,434 (824)	12,338 (846)	14,913 (960)	9,705 (1,071)	7,455 (817)
3,0	640 (354)	0 (452)	0 (288)	1,988 (562)	3,068 (383)
2,2	7,078 (440)	14,146 (445)	6,632 (428)	10,127 (659)	10,535 (524)
1,3;3,1	8,212 (450)	14,124 (465)	10,045 (429)	12,432 (713)	10,083 (527)
4,0	5,474 (548)	5,704 (515)	5,423 (413)	4,291 (726)	9,226 (567)
3,2;2,3	5,106 (548)	7,714 (563)	5,054 (435)	8,983 (872)	3,032 (605)
4,1;1,4	12,265 (658)	14,124 (599)	9,296 (500)	16,870 (1,025)	12,509 (819)
5,0	6,206 (642)	4,981 (584)	7,160 (511)	4,270 (888)	731 (612)
3,3	0 (634)	0 (402)	1,262 (622)	1,260 (773)	0 (577)
4,2;2,4	10,859 (745)	6,246 (656)	8,171 (546)	5,962 (1,071)	5,178 (872)
5,1;1,5	4,176 (963)	3,111 (690)	2,246 (437)	2,952 (801)	1,008 (670)
6,0	0 (667)	0 (471)	0 (477)	0 (577)	0 (432)
4,3;3,4	34,885 (1,304)	24,456 (1,730)	30,158 (1,946)	26,013 (2,290)	13,105 (1,400)
5,2;2,5	8,665 (720)	7,885 (819)	6,569 (708)	5,802 (851)	7,195 (719)
6,1;1,6	4,365 (821)	4,235 (871)	2,915 (757)	3,082 (902)	1,342 (586)
4,4	0 (768)	0 (739)	1,195 (654)	0 (675)	0 (506)
7,0;5,3;3,5	6,942 (829)	9,750 (990)	6,970 (823)	4,656 (972)	2,442 (691)
6,2;2,6	1,022 (733)	2,070 (849)	0 (608)	0 (731)	372 (520)
7,1;1,7	6,023 (1,017)	6,439 (1,140)	4,335 (941)	3,965 (1,095)	2,661 (768)

the helices or errors in the assignment of helical regions in the sequence. There is currently insufficient structural information available to estimate analytically the impact of these deficiencies; however, the fact that models have been found that give *R* values close to the expected minimum *R* value for several data pairs suggests that these factors might not be prohibitory. Furthermore, the use of a number of different, independent derivatives provides assurance that false models will be rejected.

Model Calculations

Amino acid species that are distributed unevenly throughout the helices are likely to be the most valuable in identifying helices. Initial efforts to make helix assignments concentrated on trying to locate heavily labeled helices that would be likely to dominate the difference data. Table III shows the distribution of the labeled amino acids throughout the helices, while Fig. 3 shows the location of these residues for each helix in projection. From Table II and Fig. 3 note that helix F contains seven valine residues (40% of the total number of valines distributed throughout the helices) and that these residues are largely clustered on one face of the helix. The next most valine-rich helices are D and G, containing three valine residues each, and, in the case of helix G, the three valines are clustered near each other. The distribution of a given kind of amino acid along a helix becomes important in the model building when the helices are tilted (as for helices in positions 1, 2, 3, and 4) or if a helix is significantly distorted from an ideal alpha helix.

A series of model tests has been executed, the flow of which is described in Fig. 4. All 7! = 5,040 assignment models were tested against the native/D Val difference data, allowing threefold rotations of each helix. Two sets of threefold rotations, 60° out of phase with each other, were examined. A rotation of one helix by 120° corresponds to an average movement of ~10 Å of an amino acid side chain. While the resolution of the data is 7 Å, this coarse rotation search should be sufficient to locate the heavily labeled F helix. It is computationally prohibitive to test all possible assignment models with a finer rotation search. From this threefold rotation search only 0.002% of the total number of models tested (2.2×10^7) gave *R* values in

TABLE II
INCORPORATION OF LABELED AMINO ACIDS

Derivative	Percent incorporation	Method of determination
D Val	100	radioactive analysis
D Ile	74	GC/mass spectroscopy
D Leu	77	GC/mass spectroscopy
D Phe	60	^{1}H NMR spectroscopy

TABLE III
DISTRIBUTION OF AMINO ACIDS IN SEQUENCE

	A	B	C	D	E	F	G	Links
Val	1	1	1	3	2	7	3	3
Phe	1	2	1	0	4	0	2	3
Leu	6	4	8	4	3	3	6	2
Ile	1	2	1	3	2	2	2	2

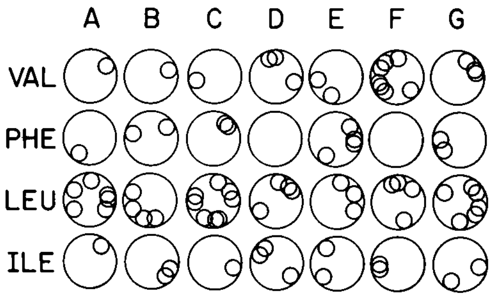

FIGURE 3 Projections of the seven helices showing the disposition of the side chains. Each helix is diagrammed in end on projection, and is shown in the same rotational position for each of the deuterated amino acid distributions for comparison.

the range 26–40%. Table IV summarizes the information on the positions of helices F and G from these models. All the models with R values <40% placed helix F in density positions 1, 2, 3, or 4 with consistent respective rotation assignments. Of these four positions, position 3 occurred by far the most frequently. When assignment of a heavily labeled helix to a position leads to a large number of models with small R's the probability that this helix is correctly assigned is high. Helix G is also favored in positions 1, 2, 3, or 4 with position 4 occurring the most frequently.

The next step in our analysis was to sequentially assign helix F to density positions 1, 2, 3, and 4 using the best rotation assignments indicated in the previous calculation, and search all possible assignment models, with sixfold rotations allowed for the remaining helices. For this calculation, helices A, B, and C, containing one valine each, were considered equivalent and were not permuted with each other, though they were free to rotate. A total of 696 assignment models were found to give good agreement ($19 < R < 30$) with the native/D Val differences. Again helices F and G were most frequently placed in positions 3 and 4, respectively (Table IV).

Calculations were also done using the native/D Phe difference data. From Table III it is seen that helices D and F contain no phenylalanine. This means that these helices need not be rotated or permuted with each other in testing models against native/D Phe differences, making feasible a complete search of all possible assignment models with sixfold rotations on the phenylalanine containing helices. Such a calculation was done. The minimum R value obtained was 40%, and 358 assignment models were found with R values <50%. Note also from Table III and Fig. 3 that helix E contains four phenylalanine residues, three of which are clustered together. Most of the models that gave

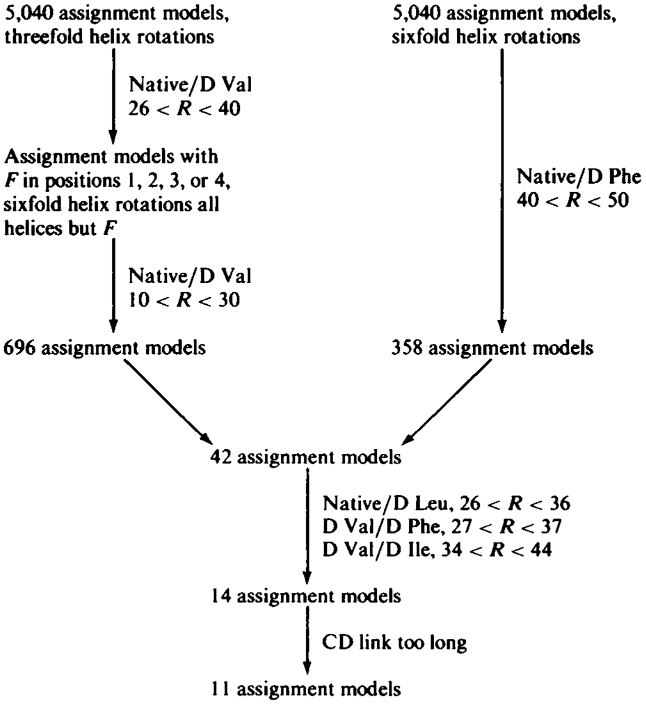

FIGURE 4 Flow diagram of model building calculations.

TABLE IV
POSITIONS OF HELICES F AND G FROM NATIVE/D VAL CALCULATIONS

Helix density	Calculation 1* N§ $26 < R < 40$		Calculation 2‡ N§ $19 < R < 30$	
	F	G	F	G
1	76	55	205	228
2	50	88	376	192
3	168	36	1,175	133
4	38	150	33	1,189
5	0	0	—	2
6	0	0	—	0
7	0	3	—	45

*5,040 assignment models tested with three-fold helix rotations (2.2×10^7 models).

‡Assignment models with F in density positions 1, 2, 3, and 4, tested with sixfold helix rotations (1.3×10^8 models).

§Number of models with R values in the range indicated and with helices F and G in the positions specified.

TABLE V
POSITIONS OF HELIX E FROM NATIVE/D PHE CALCULATIONS

Density position	N* $40 < R < 50$
1	28
2	286
3	48
4	80
5	2
6	10
7	142

*Number of models tested having R values in the range indicated with helix E in the positions specified. All 5,040 assignment models were tested with sixfold helix rotations (1.41×10^9 models).

good agreement with the native/D Phe differences placed helix E in density positions 2 or 7 (Table V). This can be understood if the phenylalanine side chains are disposed toward the space between these density positions. The comparatively high minimum R value found for the native/D Phe differences is not unexpected. Phenylalanine has an extended side chain, and rotational freedom about the C_α—C_β bond can move the center of the ring system ~5 Å. In addition, 25% of the total number of phenylalanine residues are in the linking regions and therefore not included in the model.

TABLE VI

R VALUES FOR ASSIGNMENT MODELS GIVING GOOD AGREEMENT WITH NATIVE/D VAL AND NATIVE/D PHE DIFFERENCES

Model	Native D Val	Native D Phe	Native D Leu	D Val D Phe	D Val D Ile	Bad CD link*
AEFGDBC	20	40	27	27	40	
AEFGDCB	20	44	27	32	34	
ABFGDEC	19	49	26	28	44	
BEFGDAC	20	42	35	30	42	
ABFGDCE	21	47	28	30	40	
DEFGABC	22	46	32	28	40	
DEFGBAC	22	50	31	28	40	
DBFGACE	24	49	32	32	40	
AEFGBDC	25	48	29	34	42	
DEFGACB	22	50	31	37	41	
BAFGDCE	21	50	29	33	37	
DEFGBCA	22	49	37‡	31	41	
ACFGDBE	21	47	27	31	45‡	
BEFGADC	25	44	33	33	45‡	
BEFGCDA	25	50	36	33	46‡	
EBFGACD	20	48	33	32	48‡	
AEFGBCD	22	48	32	31	47‡	
BEFGDCA	20	47	>40‡	31	43	
AGFDCBE	23	49	30	38‡	>50‡	
AGFEBCD	23	46	37‡	38‡	>50‡	
AEFBDCG	29	47	29	>40‡	>50‡	
AEFCDBG	29	45	32	>40‡	>50‡	
BEFCDAG	29	46	38‡	>40‡	>50‡	
CAFGDBE	21	48	34	29	42	X
CEFGDBA	20	49	33	29	37	X
CEFGDAB	20	48	34	29	44	X
CBFGDAE	21	50	>40‡	27	41	X
CBFGDEA	19	50	37‡	27	47‡	X
DEFGCBA	22	47	>40‡	29	40	X
CEFADBG	29	48	33	>40‡	>50‡	X
FCEGBDA	29	49	>40‡	>50‡	>50‡	
FGEABDC	24	50	39‡	>40‡	>50‡	
FGBEACD	27	49	38‡	>40‡	>50‡	
FEDGCBA	29	47	>40‡	32	>50‡	
FEDGBAC	29	50	39‡	31	>50‡	
FBDGACE	30	49	36	38‡	>50‡	
FEDGACB	29	47	35	33	>50‡	
FEDGABC	29	46	35	29	>50‡	
BCGFDAE	29	46	39‡	>40‡	49‡	
BAGFDCE	29	49	>40‡	>40‡	46‡	
ABGFDCE	29	49	32	>40‡	>50‡	
ACGFDBE	29	49	37‡	>40‡	>50‡	
$R_{\min}$§	19	40	26	27	34	

*Bad CD link indicated by X for models with helices C and D in positions 1 and 5 or 5 and 1, respectively.

‡Indicates the *R* value found for the assignment model is more than 10% higher than the minimum *R* value found for the data pair.

§Minimum *R* value observed for the data pair.

Only 42 assignment models had *R* values that fell in the lowest 10% of the *R* distributions for both the native/D Val and native/D Phe differences (Table VI). These models were tested against native/D Leu, native/D Ile, D Val/D Phe, and D Val/D Ile differences. It was found that three of these data pairs discriminate among the 42 models. Requiring that the *R* values fall in the lowest 10% for these data pairs eliminated all but 14 assignment models. Although the *R* values calculated for the native/D Ile differences were low, these data were not useful in discriminating between models, probably as a result of the relatively even distribution of isoleucine among the helices (Table III) and the spread distribution of isoleucine within some helices (e.g., helix G, Fig. 1). When paired with other derivatives, however, D Ile data proved useful.

An additional constraint is provided by the lengths of the linking regions between helices. In the case of helices C and D, the distance between the ends of the helices in some models exceeded the maximum possible length attributable to the linking amino acids by a factor of 2. Three of the 14 models were rejected using this constraint, leaving 11 assignment models.

The 11 assignment models were then tested using the four remaining combinations of the five data sets (Table VII). In the model building analysis, any pair of data sets may be used as the basis of model selection. It is expected that the use of two deuterated derivative data sets will usually result in a larger minimum *R* value than the comparison of a derivative with the native data, since a stronger requirement is placed on the model by the double derivative comparison when two sets of modified sites are simultaneously evaluated. In general, our observations

TABLE VII

R VALUES FOR 11 BEST ASSIGNMENT MODELS

	Model	D Val/ D Leu	D Phe/ D Ile	D Phe/ D Leu	D Ile/ D Leu
1	AEFGDBC	51	40	43	44
2	AEFGDCB	50	48*	40	45
3	ABFGDEC	50	50*	45	46
4	BEFGDAC	53	35	43	44
5	ABFGDCE	50	39	40	47
6	DEFGABC	52	43	42	50*
7	DEFGBAC	55	42	43	45
8	DBFGACE	54	36	37	37
9	AEFGBDC	49	47*	42	37
10	DEFGACB	51	37	35	40
11	BAFGDCE	61*	42	41	48*
	$R_{\min}$‡	49	35	35	37

*Indicates the *R* value found for the assignment models is more than 10% higher than the minimum *R* value found for the data pair.

‡Minimum *R* value found for the data pair.

support this view. The "best" models in Table VI were found to be in good agreement with the combined data, as they should be (Table VII). Models that were marginally excluded by the previous criteria (models 12–17, Table VII) were strongly rejected by the remaining data pairs. The 11 best models fell in the lowest 15% of R in all cases, and were mostly in the lowest 10%. Note that all surviving models placed helices F and G in density positions 3 and 4. As a check that the minimum R values observed for each data pair was indeed a global minimum, R values were calculated for all possible assignment models with helices F and G in density positions 3 and 4, with fixed rotations as indicated in previous calculations. The minimum R values found in these searches did not differ from the minimum R values found for the 11 models by more than a few percent. It is important to recognize that the mean R value for each comparison was in excess of 100%, so that models falling in the lowest 10% are selected by a strong criterion (18). Also, the observation that models among the 42 selected using three of the five data sets were compatible with the remaining two and with all combinations gives strong support to the correctness of the selection.

The 11 models selected all place helices F and G in density positions 3 and 4 in the structure, a result strongly indicated by the initial native/D Val calculations. Furthermore, the rotations assigned to these helices are consistent between independent data pairs. It is of particular interest that for all the models, helix G is oriented so that lysine 216, which binds the retinal chromophore (19), is directed toward the space in between density positions 4 and 5. A more detailed examination of helix orientations must await more specific helix assignments for helices A–E. These helices have not been uniquely located in this analysis though there are strong indications. Helix C is consistently placed in positions 6 or 7; helix E is placed in positions 2 or 7, with one exception; helix D is placed in positions 1 or 5, with one exception; helices A and B are less consistently placed. It is interesting to note that 10 of the models place helix D, which has a relatively low electron scattering cross section, in the weak density positions (1 and 5) of the electron microscope map (5).

The assignment of helices F and G to density positions 3 and 4 contradicts all previously proposed assignment models for bacteriorhodopsin (4–6) as well as the location of the retinal chromophore suggested by King et al. (8) and the location of the C terminus reported by Wallace and Henderson (9).

Using a constrained density map modification and refinement procedure, Agard and Stroud (6) proposed a set of linking regions in bacteriorhodopsin.They predict that helices in positions 3 and 4 in the structural map would be linked on the cytoplasmic side of the membrane. Our assignment would place the link region on the external side.

King et al. (8) have reported a location for the retinal chromophore using a difference Fourier analysis of two-dimensional neutron diffraction data from native purple membranes and purple membranes reconstituted with deuterated retinal. The location of the retinal on lysine 216 (19) would place it next to the G helix (20), which our analysis placed in position 4. This location is incompatible with the location proposed by King et al., which could be reconciled with helix G being in positions 1, 2, 7, or 6. Similarly, the location of the C terminal segment of the polypeptide should be closely related to the position of helix G. The location reported by Wallace and Henderson (9) is incompatible with our location for G. They calculated difference Fourier maps using electron diffraction data on native purple membranes and purple membranes with the carboxyl terminus removed. Their interpretation of these maps was to place helix G in density position 2.

Finally, among the 41 models proposed by Engelman et al. (5) there are only two with helices F and G in positions 3 and 4, and these models are not among the 11 selected by our present analysis. In our assignment of helices in the amino acid sequence (15), we have not placed lysine 171 and arginine 174 in the nonpolar regions of the bilayer. This assignment is supported by recent crosslinking results (20). Thus, two of the groups previously assumed to be involved in ion pairing are removed from consideration, and it is not surprising that agreement between our neutron analysis and this earlier approach is not found.

The conclusion that is most firmly established by our analysis, that helices F and G are in positions 3 and 4, casts doubt on many earlier ideas for the structural organization of bacteriorhodopsin and establishes important constraints on future efforts to understand the chemical organization of the molecule. The use of additional derivatives and refinement procedures should permit a unique choice of model in further applications of the neutron diffraction methods we have developed.

We are grateful for the skilled technical assistance of B. Gillette, A. Perlo, R. Rice, and S. Wilson. We thank M. Dumont, R. Henderson, and D. Agard for helpful discussions. Special thanks are given to D. LeMaster for invaluable discussions and advice concerning the isolation of deuterated amino acids.

We thank the National Institutes of Health (GM 29243, GM 22778) and the National Science Foundation (PCM 78-10361) for support, and the Institut Laue-Langevin for the use of their excellent facilities.

Received for publication 17 November 1982 and in final form 7 February 1983.

REFERENCES

1. Stoeckenius, W., R. H. Lozier, and R. A. Bogomolini. 1979. Bacteriorhodopsin and the purple membrane of *Halobacteria*. *Biochim. Biophys. Acta.* 505:215–278.
2. Henderson, R., and P. N. T. Unwin. 1975. Three-dimensional model of purple membrane obtained by electron microscopy. *Nature* (*Lond.*). 257:28–32.
3. Khorana, H. G., G. E. Gerber, W. C. Herlihy, C. P. Gray, R. J. Anderegg, K. Nihei, and K. Biemann. 1979. Amino acid

sequence of bacteriorhodopsin. *Proc. Natl. Acad. Sci. USA.* 76:5046–5050.

4. Ovchinnikov, Yu. A., N. G. Abdulaev, M. Yu. Feigina, A. V. Kisllev, and N. A. Lobanov. 1979. The structural basis of the functioning of bacteriorhodopsin: an overview. *FEBS* (*Fed. Eur. Biochem. Soc.*) *Lett.* 100:219–224.
5. Engelman, D. M., R. Henderson, A. MacLaughlin, and B. A. Wallace. 1980. The path of the polypeptide in bacteriorhodopsin. *Proc. Natl. Acad. Sci. USA.* 77:2023–2027.
6. Agard, D. A., and R. M. Stroud. 1981. Linking regions between helices in bacteriorhodopsin revealed. *Biophys. J.* 37:589–602.
7. Engelman, D. M., and G. Zaccai. 1980. Bacteriorhodopsin is an inside-out protein. *Proc. Natl. Acad. Sci. USA.* 77:5894–5898.
8. King, G. I., W. Stoeckenius, H. L. Crespi, and B. P. Schoenborn. 1979. The location of retinal in the purple membrane profile by neutron diffraction. *J. Mol. Biol.* 130:395–404.
9. Wallace, B. A., and R. Henderson. 1982. Location of carboxyl terminus of bacteriorhodopsin in purple membrane. *Biophys. J.* 39:233–239.
10. Oesterhelt, D., and W. Stoeckenius. 1971. A rhodopsin-like protein from the purple membrane of *Halobacterium halobium. Nature* (*Lond.*). 233:149–154.
11. LeMaster, D., and F. M. Richards. 1982. Preparative-scale isolation of isotopically labeled amino acids. *Anal. Biochem.* 122:238–247.
12. Keefer, L. M., and R. A. Bradshaw. 1977. Structural studies on *Halobacterium halobium* bacteriorhodopsin. *Fed. Proc.* 36:1799–1804.
13. Gehrke, C. W., H. Nakamoto, and R. W. Zumwati. 1969. Gas-liquid chromotography of protein amino acid trimethylsilyl derivatives. *J. Chromatog.* 45:24–51.
14. Zaccai, G., and D. J. Gilmore. 1979. Areas of hydration in purple membrane of *Halobacterium halobium:* a neutron diffraction study. *J. Mol. Biol.* 132:181–191.
15. Rogan, P. K., and G. Zaccai. 1981. Hydration in purple membrane as a function of relative humidity. *J. Mol. Biol.* 145:281–283.
16. Engelman, D. M., A. Goldman, and T. A. Steitz. 1981. The identification of helical segments in the polypeptide chain of bacteriorhodopsin. *Methods. Enzymol.* 88:81–88.
17. Diamond, R. 1979. Bilder User's Guide. MRC Laboratory of Molecular Biology, Cambridge, England. 20–24.
18. Trewhella, J., E. Gogol, G. Zaccai, and D. M. Engelman. 1983. Neutron diffraction studies of bacteriorhodopsin. Brookhaven Symposium in Biology. Vol. 27. In press.
19. Lemke, H. D., and D. Oesterhelt. 1981. Lysine 216 is a binding site of the retinyl moiety in bacteriorhodopsin. *FEBS* (*Fed. Eur. Biochem. Soc.*) *Lett.* 128:255–260.
20. Bayley, H., K.-S. Huang, R. Radhakrishnan, A. H. Ross, Y. Takagaki, H. G. Khorana. 1981. Site of attachment of retinal in bacteriorhodpsin. *Proc. Natl. Acad. Sci. USA.* 78:2225–2229.
21. Huang, K. S., H. Bayley, and H. G. Khorana. 1981. Orientation of retinal in bacteriorhodopsin. *Fed. Proc.* 40:1659.