The EMBO Journal vol.6 no.1 pp.265–273, 1987

The halo-opsin gene. II. Sequence, primary structure of halorhodopsin and comparison with bacteriorhodopsin

A.Blanck and D.Oesterhelt

Max-Planck-Institut für Biochemie, D-8033 Martinsried, FRG

Communicated by D.Oesterhelt

The gene for the protein moiety of the light-driven chloride pump halorhodopsin (HR), *hop* gene, was sequenced and the primary structure of the protein derived thereof. The gene has a GC content of 67% and codes for 274 amino acids. A promoter structure, resembling that of the halobacterial 16S rRNA genes, is present and both a terminating stem and a loop sequence is found downstream of the TGA stop codon. A ribosomal binding site is located within the translated region. The HR protein moiety is processed at the amino terminus, as well as the carboxy terminus, yielding a dominant species of calculated M_r 26 961. Seven transmembrane helical parts of the protein are defined by hydropathy and acrophilicity calculations. Comparision with the bacteriorhodopsin (BR) structure reveals a conservation of 36% of amino acid residues in the transmembrane part and 19% in the connecting loops at both surfaces. The most conspicuous conserved amino acids are the retinal-binding Lys residue, four Trp residues (eventually interacting with retinal), two Asp residues (providing possibly the negative charge environment of retinal) and three Pro residues of unknown function. No significant homology with the opsins of eucaryotes was found. Helical wheel analysis shows that HR is an inside-out protein with the majority of conserved amino acid residues inside the circle of the seven transmembrane helices. It is postulated that the intrahelical spaces, which could be gated by the retinal moiety, are the physical entities for translocation of protons in BR and chloride ions in HR. Retinal, by its *cis–trans* isomerization, serves as a switch connecting the ion-specific binding sites in both proteins.

Key words: halorhodopsin/chloride pump/gene sequence/primary structure/amino acid conservation

Introduction

In the preceding paper (Hegemann *et al.*, 1987) the isolation of the gene coding for the light-driven chloride pump halorhodpsin (HR) was described. The nucleotide sequence reported in this contribution allows, together with pre-existing amino acid sequence data, the primary structure of HR to be derived. On the basis of the similarity with its proton translocating counterpart, bacteriorhodopsin (BR), predictions regarding the secondary structure of HR and a general mechanism for ion translocation in retinal proteins should be possible.

Results

The gene sequence

The halo-opsin gene (*hop*) is located on a 36-kb fragment of the halobacterial genome and was isolated from a cosmid library of a *Sau*3AI partial digest (Hegemann *et al.*, 1987). Figure 1 shows the restriction map of the cosmid pAB H47 containing this fragment as insert and, in more detail, the part which was used for the sequencing of the *hop* gene. Smaller segments were subcloned into the phage vector M13mp8 and the areas, indicated by arrows from left to right in Figure 1, were sequenced by the dideoxy-method of Sanger (Sanger *et al.*, 1977). The subclone containing the *Acc*I fragment was too large to be sequenced in a single run. Therefore, two additional synthetic oligodeoxynucleotides were used as primers. For unknown reasons all but one fragment (indicated by two arrow heads) were inserted into M13 exclusively in one direction and this provided the sequence of only one strand. The second strand was therefore sequenced separately with the help of the exonuclease technique (Henikoff, 1984) (as indicated by the arrows pointing to the left in Figure 1).

In Figure 2 the results of all sequencing experiments are summarized. On the 1291-bp segment shown, an open reading frame (ORF) starts at position 283 with an ATG codon and ends at position 1105 with an TGA codon. The ORF codes for 274 amino acids corresponding to a protein with a mol. wt. of 28 874 daltons. Protein chemical studies had suggested earlier that the N-terminus of the majority of the HR molecules is blocked (Schegk *et al.*, 1986), or heterogeneous, but ~25% of the molecules start at position 22 with the sequence A-V-R-E-N-A-L-L yielding a protein of M_r 26 961. Minor quantities of peptides with N-termini upstream of this sequence were found and thus it remains an open question whether the *hop* gene product as isolated from cell membranes has a defined N-terminus at all (M.Kehl, unpublished results). In any case, the message transcribed from the gene starts six bases upstream of the ATG codon and thereby confirms the assumption that translation starts with the Met at nucleotide position 283 (Blanck and Oesterhelt, in preparation). The stretches underlined in Figure 2 were confirmed by sequencing of either pure peptides, or peptide mixtures. These partial sequences cover the ORF in a quite regular distribution and thus allow the firm conclusion that the translation of the ORF indeed represents the primary structure of HR. This was further supported by the synthesis of a peptide from amino acid positions 85–96 and production of polyclonal antibodies against this peptide that cross-reacted with native HR (May, 1986).

The *hop* gene has a GC content of 67% and a codon usage similar to that of the bacterio-opsin (bop) and 'BR-related protein' (brp), as summarized in Table I. The predominance of C and G at the third position is evident in all three cases, e.g. for Leu, but especially for the *hop* gene, which has not a single codon for Pro, Thr or Ala with an A or T at the third position.

Promoter, terminator and ribosomal binding site

From the archaebacterial genes so far sequenced some promotor-specific sequences have become apparent, especially from the multiple promoter sites on the halobacterial 16S rRNA gene (Mankin *et al.*, 1984; Dennis, 1985) which are characterized by four motifs. In this 16S rRNA a 23-mer (motif 1) is followed at a distance of four nucleotides by the sequence TTCGA (motif 2) (repeated after five nucleotides). After 18 nucleotides the

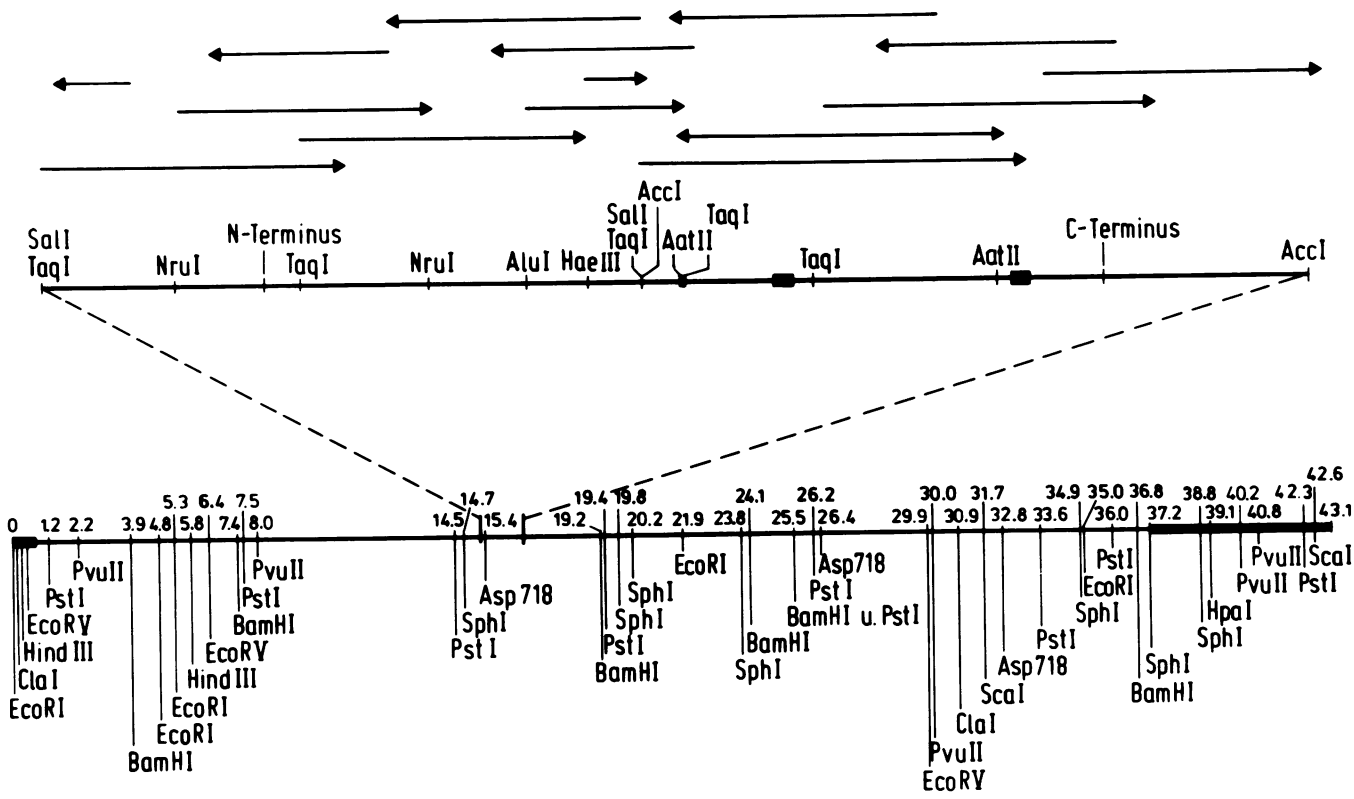

Fig. 1. Restriction map of pAB H47 established with the endonucleases *Asp*718, *Bam*HI, *Cla*I, *Eco*RI, *Eco*RV, *Hin*dIII, *Hpa*I, *Pvu*II, *Pst*I, *Sca*I, *Sph*I. The vector part is shown as a thick line and the insert as a thin one. The numbers are the distances in kilobases of respective restriction sites from the *Eco*RI site at 0. The segment carrying the *hop* gene is shown enlarged above the insert and those restriction sites are indicated that were used for subcloning and sequencing. The arrows indicate the sequenced fragments and the thick lines the two synthetic primers (positions 785–803 and 1008–1035 in the sequence of Figure 2) used for sequencing of the *Acc*I fragment. The following fragments were subcloned in M13: *Sal*I fragment position 1–624 (M13SS9), *Taq*I fragment position 312–844 (M13AT6, described in Hegemann *et al.*, 1987), *Alu*I/*Taq*I fragment position 510–668 (M13SA23), *Hae*III/*Taq*I fragment position 575–625 (M13SH11), *Acc*I fragment position 626– ~1315 (M13AA6), *Aat*II fragment position 669–998 (M13SA 3 and 15) and *Nru*I fragment position 141–409 (M13SN3).

Table I. Comparison of codon usage of the three halobacterial genes of known sequence and coding for proteins (not yet unequivocal for *brp*)

		hop	*bop*	*brp*			*hop*	*bop*	*brp*			*hop*	*bop*	*brp*			*hop*	*bop*	*brp*
UUU	Phe	20	0	5	UCU	Ser	0	0	0	UAU	Tyr	29	27	8	UGU	Cys	0	0	0
UUC		80	100	95	UCC		17	21	9	UAC		71	73	92	UGC		100	0	0
UUA	Leu	0	5	0	UCA		8	0	5	UAA	OC	0	0	0	UGA	OP	100	100	100
UUG		10	15	13	UCG		42	36	57	UAG	AM	0	0	0	UGG	Trp	100	100	100
CUU	Leu	5	5	9	CCU	Pro	0	0	0	CAU	His	0	0	40	CGU	Arg	0	43	0
CUC		32	31	37	CCC		43	25	31	CAC		100	0	60	CGC		36	43	33
CUA		3	5	0	CCA		0	25	19	CAA	Gln	25	0	0	CGA		9	0	11
CUG		50	39	41	CCG		57	50	50	CAG		75	100	100	CGG		46	14	50
AUU	Ile	0	0	8	ACU	Thr	0	0	14	AAU	Asn	0	0	33	AGU	Ser	21	7	0
AUC		93	100	92	ACC		44	37	29	AAC		100	100	67	AGC		12	36	29
AUA		7	0	0	ACA		0	10	0	AAA	Lys	0	43	50	AGA	Arg	9	0	0
AUG	Met	100	100	100	ACG		56	53	57	AAG		100	57	50	AGG		0	0	6
GUU	Val	0	13	5	GCU	Ala	0	3	2	GAU	Asp	22	20	0	GGU	Gly	4	15	17
GUC		26	35	59	GCC		18	27	47	GAC		78	80	100	GGC		29	39	52
GUA		3	13	0	GCA		0	17	9	GAA	Glu	0	36	0	GGA		17	15	7
GUG		71	39	36	GCG		82	53	42	GAG		100	64	100	GGG		50	31	24

The values are expressed as percent of amino acid residues occurring. For absolute values see Table II.

polymerase binding site AAGTAA (motif 3) occurs separated by another 20 nucleotides from the sequence TGCGAACG (motif 4) with the start of the transcript at the second A. In Figure 2 upstream sequences of the *hop* gene are marked in italics, which show 67% homology (bold print) to the motifs 1, 2 and 4 of the promoter sequence arrangement of the *Halobacterium halobium* and the *H. cutirubrum* 16S rRNA. However, motif 3, the polymerase binding site, does not occur in place. Instead this

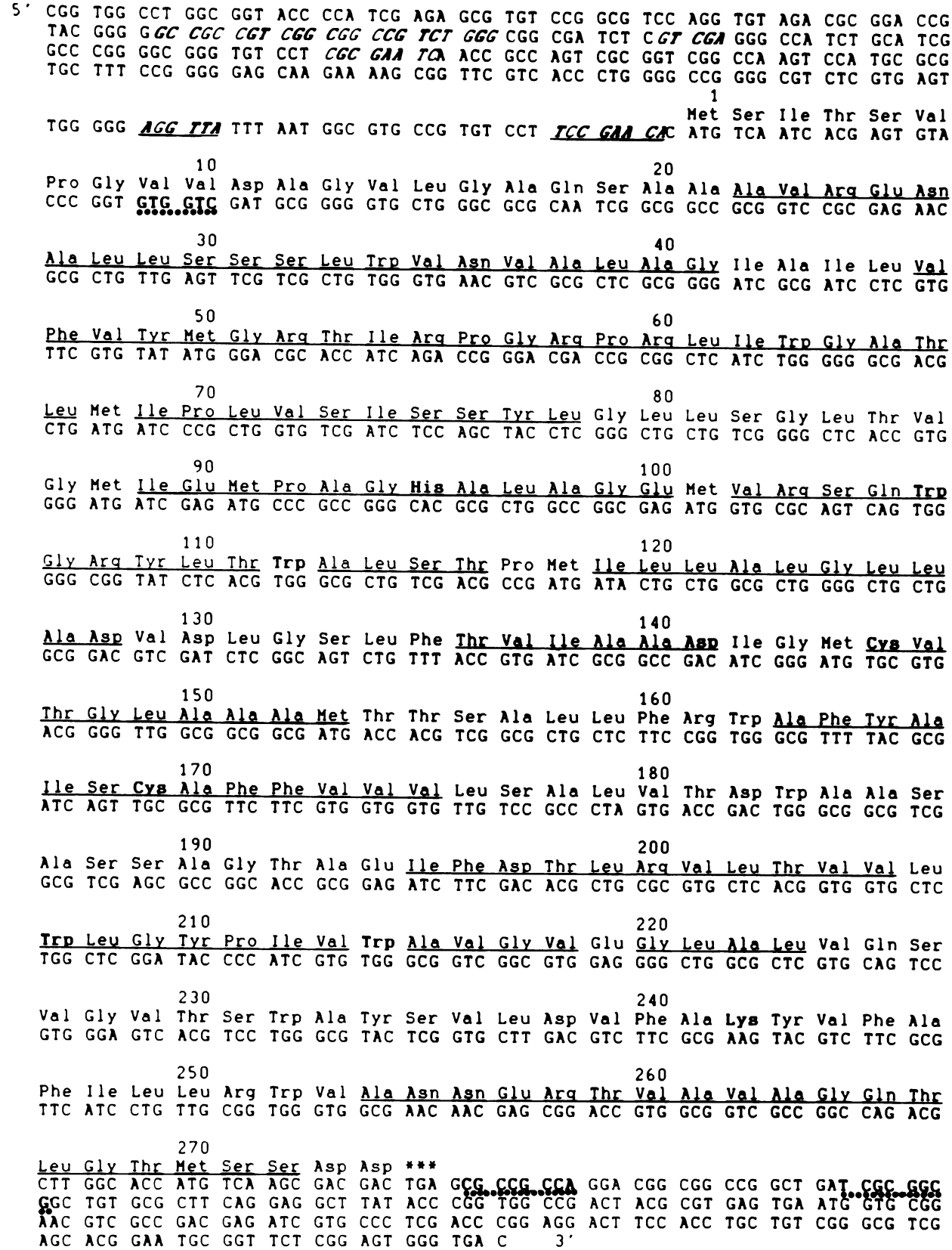

Fig. 2. Gene sequence of *hop* with its flanking upstream and downstream sequences. An ORF starts at position 283 with a ATG codon and ends at position 1105 with a TGA stop codon. The underlined stretches of the amino acid sequence derived from the nucleotide sequence were confirmed by protein chemical experiments (F.Lottspeich, M.Kehl and S.Schegk, personal communication), or mass spectroscopy, after chemical cleavage of the gene product and h.p.l.c. purification of peptide fragments (see Hegemann *et al.*, 1987). For promoter structural motifs (italics, bold print and underlined), ribosomal binding site (pointed area) and terminator sequence (pointed areas) see text and Figure 6.

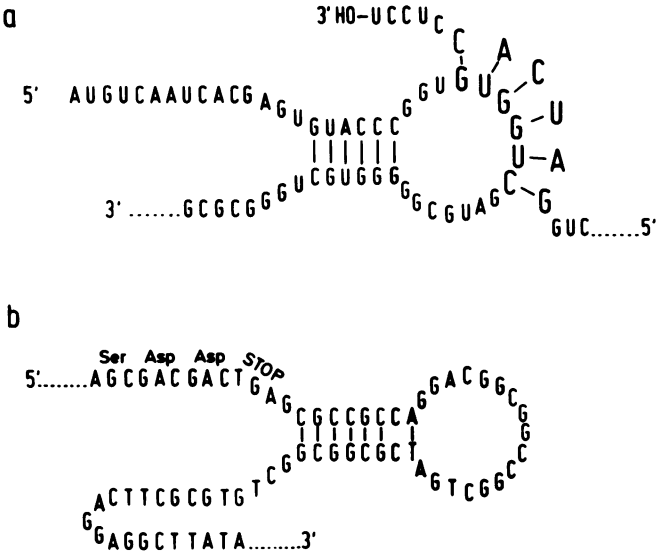

Fig. 3. Potential ribosomal binding site and terminator sequence of the *hop* gene.

Table II. Amino acid composition and conservation in the two ion pumps BR and HR.

Amino acid	Total				Membrane			
	hop	bop	Cons. res.	% in hop	hop	bop	Cons. res.	% in hop
Phe	10	13	3	30	10	11	3	30
Leu	38	39	18	47	29	35	18	62
Ile	15	15	2	13	11	12	2	18
Met	10	10	2	20	6	6	1	17
Val	34	23	6	18	21	19	6	28
Ser	24	14	2	8	15	5	1	7
Pro	7	12	3	43	3	6	3	100
Thr	18	19	6	33	12	15	6	50
Ala	39	30	11	28	12	17	8	67
Tyr	7	11	3	43	7	10	3	43
His	1	0	0	0	0	0	0	0
Gln	4	4	0	0	1	0	0	0
Asn	4	3	0	0	1	2	0	0
Lys	1	7	1	100	1	5	1	100
Asp	9	9	4	44	5	4	2	40
Glu	6	11	2	33	1	2	0	0
Cys	2	0	0	0	2	0	0	0
Trp	10	8	4	40	10	8	4	40
Arg	11	7	4	36	4	3	3	75
Gly	24	26	8	33	10	14	4	40
	274	261						

site and the RNA start site (motif 4) occurs further downstream (underlined in Figure 2) at the expected location from S1 mapping where the 5′ end of the *hop* messenger was found shortly upstream of the ATG codon. Whether the enlarged distance between the two pairs of motifs in the *hop* gene compared to the 16S rRNA gene has functional significance or not, has yet to be established.

Figure 3a shows a loop structure at the beginning of the coding region of the *hop* gene formed by self-complementarity of the mRNA. Within the loop, base pairing with the 3′-end of the 16S rRNA over the length of six nucleotides occurs. This structure is a potential ribosomal binding site and shows similarities to that of the bop message (Betlach *et al.*, 1984, and see Discussion).

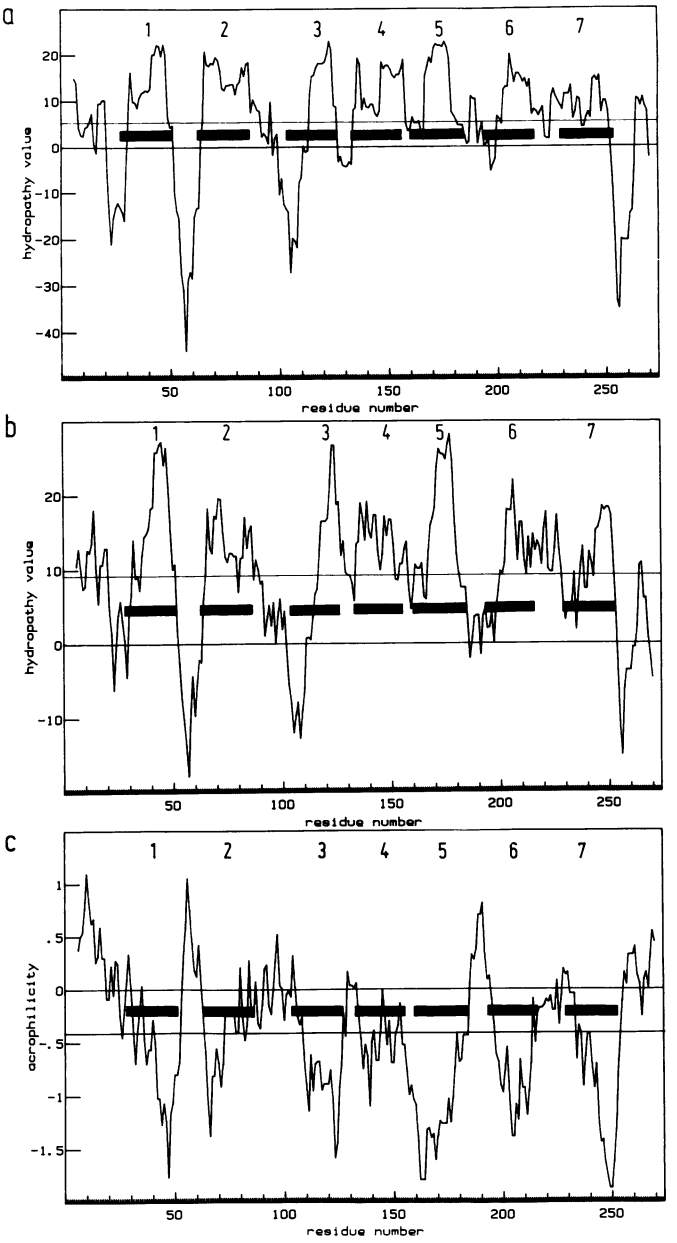

Fig. 4. Hydropathy (**a,b**) and acrophilicity (**c**) plot of HR. The hydrophobic plots were obtained with the hydrophobicity scale from Engelman *et al.* (
and Kyte and Doolittle (b). For these plots scale values of nine subseq
amino acids were summed and plotted as an average hydropathy valu
the amino acid at relative position five. For both plots the window o
was found to be optimal, whereas larger windows of up to 19 tende
further diminish the separation of helices 4/5 and 6/7. In addition to
zero line separating the hydrophobic areas (helices) 1–7 from hydrop
parts (connections) a second separation line is calculated as the mean of
274 hydropathy values. This line defines better the helical parts shown as
bars and serves as the basis for the secondary structural model in Figure 5.
The acrophilicity plot (**c**) was obtained using the scale of Hopp (1985) and
indicates, in its positive maxima, the connections between helices 1 and 7.
The line additional to the zero line represents the average acrophilic value
of all amino acids.

Figure 3b shows the sequence downstream of the translational stop codon TGA in the arrangement typical for terminator structures. A GC-rich self-complementary region (stem and loop) is followed by a relatively AT-rich (50%) sequence.

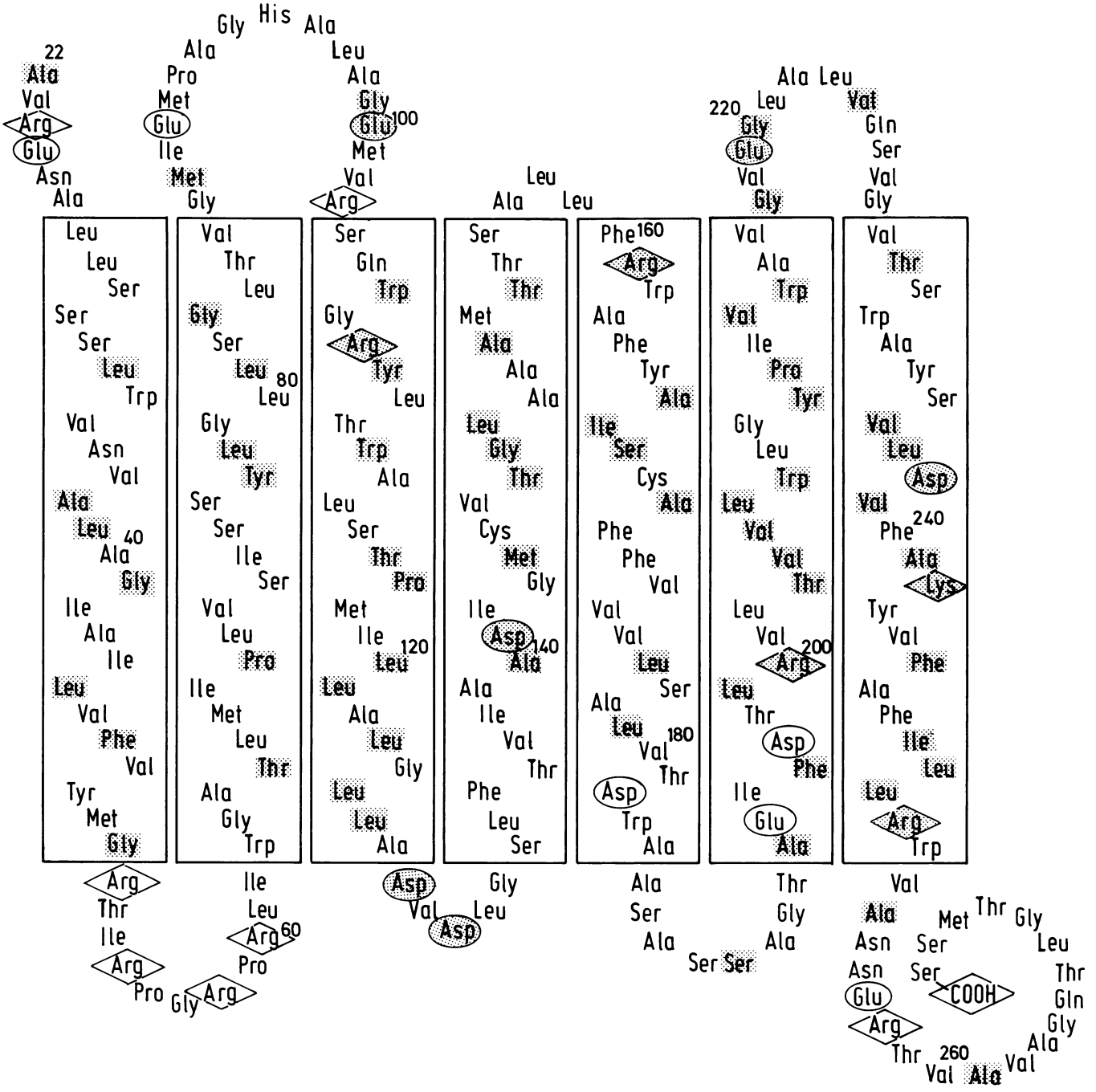

Fig. 5. Secondary structural model of HR and sequence comparison with BR. Circles indicate acidic residues, diamonds basic residues (not including His) and shaded areas show residues which occur at identical positions in BR and HR. For gaps see Figure 6. The main sequence of the isolated protein is shown. By shift of the baselines in Figure 4 the assumed length of the helices can be varied (see also Engelman *et al.*, 1986).

The protein

The amino acid composition of the primary translation product halo-opsin as defined by the ORF (Figure 2) is given in Table II. The molecule has only one lysine at position 242, which must correspond to the retinal-binding site, because by Resonance Raman spectroscopy it has previously demonstrated that retinal occurs in HR as a protonated Schiff base (Smith *et al.*, 1984; Alshuth *et al.*, 1985). Histidine occurs only once in the sequence, whereas all other positive charges are contributed by 11 arginine residues. Another unexpected result is the occurrence of two cysteine residues at position 145 and 169 instead of one residue per mole HR found by amino acid analysis and titration of SH groups (Ariki and Lanyi, 1984; Schegk *et al.*, 1986 and Discussion). The 10 methionine residues, used as cleavage sites for cyanogen bromide, are unevenly distributed in the sequence and partially explain the problems that arose earlier during peptide isolation. The N-terminal sequence of ~21 amino acids could serve as a signal sequence. It is uncertain whether an amphipathic structure, such as is typical for signal sequences, can be postulated within this region. The helical wheel plot of this region (position 1–21) gives a hydrophilic sector of Thr, Asp, Glu and Pro (80°) with a sum of −9.3 on the Kyte and Doolittle scale (see

below) contrasted by a 280° hydrophobic segment with 34.8 units. The primary translation product is apparently not only processed at the amino terminus. The carboxy terminus in the isolated protein lacks two aspartic acids (one is also missing in BR) as evidenced by protein sequencing and mass spectroscopy (M.Kehl and W.Schäfer, personal communication).

The results of c.d. and i.r. spectroscopy with halorhodopsin suggest a high helical content (Jap and Kong, 1986; Pande, Steiner and Oesterhelt, in preparation). We used two hydrophobicity scales (Kyte and Doolittle, 1982, Figure 4b; Engelman *et al.*, 1986, Figure 4a) for the calculation of a hydropathy profile shown in Figure 4 and accounted for possible salt pairing of charged residues. The results obtained by using the different hydrophobicity scales are similar, with a small difference in the position of the zero line dividing the hydrophobic membrane spanning part from the hydrophilic connections. If the average hydrophobicity is used as the line of separation, both plots become very similar. The profiles indicate the feasibility of the arrangement of the molecule in seven helical transmembrane domains. This secondary structural arrangement is typical for opsins (Ovchinnikov, 1982).

On the basis of such prediction programs the first two helices are clearly separated, followed by a third helix which has a very hydrophilic start due to Gln 105 and Arg 108. Helices 3 and 4 are not well separated in Figure 4b. A hydrophilic depression caused by Asp 141 is seen in helix 4. Helices 4 and 5, as well as helices 6 and 7, again are not separated by a pronounced hydrophilic connection. This lack of separation is overcome by an acrophilicity plot according to Hopp (1985) which locates the hydrophilic areas (positive maxima in Figure 4c) between the helices 1 to 7. Again, some improvement is attained by taking the average acrophilicity values as a basis for separation between the hydrophobic and hydrophilic parts.

The black bars in Figure 4 represent the assumed transmembrane helical parts in the model of the secondary structure of HR presented in Figure 5 which is a compromise between hydropathy and acrophilicity plot prediction, charge and glycine residue distribution, helical content of the molecule and homologies with bacteriorhodopsin. At seven helical starts or ends, glycine residues can be placed which have, due to the lack of a side chain, a greater flexibility for turns than other side chains. All seven helices have a length between 24 and 26 amino acid residues. Five positive charges on the carboxy terminal side of the protein are balanced by only four negative (2 Asp, 1 Glu and Ser-COO^-) charges. On the N-terminal side two positive charges are opposed to four negative charges in those molecules which have their N-termini blocked. Thus, a net charge difference of three units is found between the two surfaces. This might be of significance for the membrane potential-supported insertion of the molecule as suggested for the reaction center complex from the photosynthetic bacterium *Rhodopseudomonas viridis* (Michel *et al.*, 1986) and suggests, by the same reasoning, the location of the N-terminus on the outside and the C-terminus on the inside of the cell membrane.

Inside the membrane all charges compensate each other but apparent steric pairing can only be expected for Asp 238 with the retinal-bound Lys 242 and for Asp 197 with Arg 200. Possibly Glu 194 and Arg 251 could also form a salt bridge, whereas Arg 108 and Arg 161 on the N-terminal side and Asp 141 and Asp 182 on the C-terminal side are not charge-paired in any obvious way. Asp 141, existing in the middle of helix 4 as an isolated charged residue, could contribute to the negatively charged environment of the retinal molecule. It should also be noted that, given the arguments on the orientation of the molecule in the membrane, Arg 108 and 161 are located on that side of the membrane where transport-associated chloride binding is expected to occur. In the model of Figure 5 all Trp and Tyr residues are located within the membrane as are three Pro residues (70, 117 and 211) and both cysteines. Indeed SH groups only become accessible to thiol reagents upon denaturation of the chromoprotein. In the helical arrangement discussed below no intramolecular disulfide bridge can be formed between helices 4 and 5, however, an intermolecular bridge is possible, thus explaining why only one free thiol group was found.

The arginine-rich stretch from positions 52 to 60 is located in the first connecting loop on the C-terminal surface. Modification experiments indicated that 71% of Arg residues (corresponding to 8 out of 11) can be modified without major changes in the spectroscopic properties of HR (Ariki *et al.*, 1986). Indeed, in Figure 5 most of the Arg residues (7) are located in the connecting loops. Two of the remaining Arg residues are close to the membrane surface and only Arg 108 and Arg 200 are positioned below and above the retinal Schiff base. These residues could serve as anion binding sites I and II (Schobert *et al.*, 1986) and could be involved in chloride transport. No indication of a salt bridge system for chloride conduction as discussed previously (Oesterhelt *et al.*, 1986) is found. In the second connecting loop Met 91 occurs, which by peptide analysis was found to be frequently oxidized in native HR; thus indicating its location on the surface of the protein as predicted by the model. In the same connection the only His residue is located, which, by its exposure to the aqueous phase, is an ideal candidate to check the validity of the model and to establish its sideness.

Comparison of the genes and primary structures of HR and BR

Figure 6 compares two DNA fragments containing the genes for halo-opsin (upper line, position 283−1106) and bacterio-opsin (lower line, position 360−1149) aligned for maximal homology. The overall sequence homology within the ORFs amounts to 50%. The most conspicuous homology occurs around the retinal binding site where a stretch of 12 bases (positions 988−999 in HO) is identical in both genes followed, after three mismatches, by the sequence of another seven identical bases. It is indeed possible to identify in a cosmid bank clones carrying the *bop* gene with an oligodeoxynucleotide probe corresponding to the amino acid sequence positions 236−242 in HR. The entire message of either of the genes, however, does not detect specifically the respective gene-carrying clones. Four insertions (1, 2, 3, 5) in the *hop* gene occur, all being multiples of three nucleotides, but out of frame, except for the insertion 3. Two deletions out of frame occur for two amino acids (4) and for one amino acid (6) with a frameshift over a stretch of five amino acids.

Whereas the *hop* gene shares sequence homologies in the upstream region of the start codon with the repetitive promoter structures of the 16S rRNA (see above), consensus sequences of the *bop* gene with these two genes are limited to the presumed polymerase recognition site GAGTTA ~25 bp upstream of the ATG codon (marked with a broken line in Figure 6). Two putative ribosomal binding sites for the *bop* and the *brp* mRNAs were suggested (Betlach *et al.*, 1984) and two were found in the *hop* gene structure (marked with —— and ..... in Figure 6). The GTGGTC sequence shown in Figure 3 occurs 21 bp downstream of the ATG codon in the *hop* gene and is found at the same position in the *bop* mRNA as GTGG. No such consensus exists for the alternative ribosomal binding site (dotted sequence in Figure 6). Thus both mRNAs seem to share the property of having their

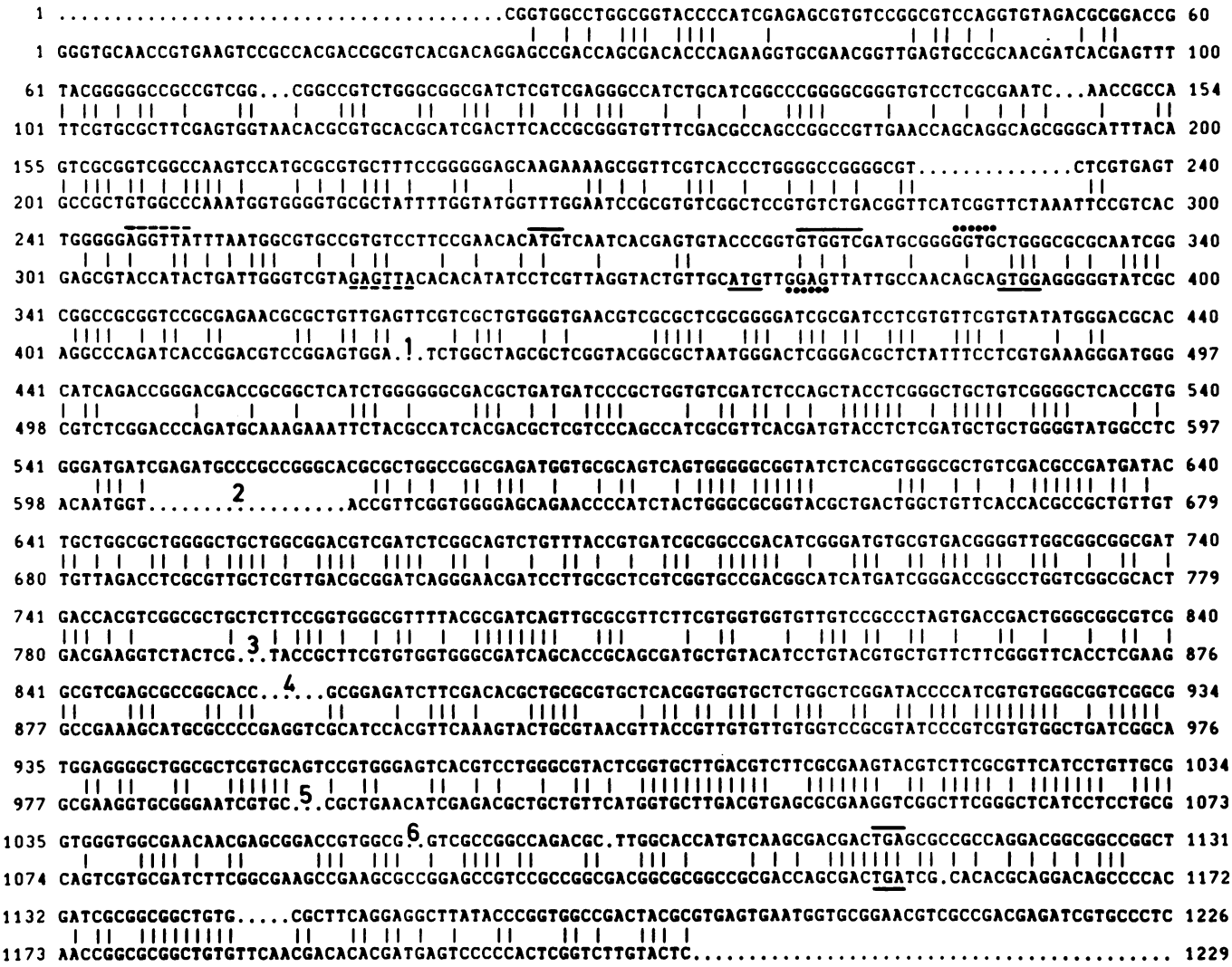

Fig. 6. Nucleotide sequence comparison of BR (lower sequence) and HR (upper sequence). The two sequences were aligned in a linear fashion by a computer program to yield maximal homologies. Deletions and insertions in the coding region are numbered 1–6 and discussed in the text. They coincide, more or less, with gaps found by manual amino acid comparison. The ATG start and TGA stop codons are under- and overlined. The potential polymerase binding sites are AGGTTA (*hop*) and GAGTTA (*bop*) and are indicated by broken lines. Out of the two potential ribosomal binding sites the dotted sequences show less consensus than the under- and overlined stretches GTGG (*bop*) and GTGGTC (*hop*).

start only a few nucleotides upstream of the start codon and the ribosomal binding site within the coding region. A similar structure comparable to the terminator sequence in the *hop* gene (as suggested in Figure 3) was also found in the *bop* gene (DasSarma *et al.*, 1984) 20 bp downstream of the TGA stop codon. A stretch of strong sequence homology between *hop* and *bop* is found within the terminator region.

The most revealing comparison of BR and HR is found on the amino acid sequence level. The residues conserved in BR are shown in the HR sequence of Figure 5 as shaded amino acids. Conservation in the transmembrane part of the molecules amounts to 36% but to only 19% in the helix connections. Among the transmembrane helices conservation varies by a factor of 2, with helices 1, 2, 4 and 5 being less homologous than helices 3, 6 and 7. In particular, helices 6 and 7 are thought to be involved in retinal–protein interaction since Lys 216 in BR and Lys 242 in HR of helix 7 are the covalent attachment points of retinal. Conservation in this area might also be expected since Resonance Raman spectroscopy revealed an almost identical protein environment of retinal in the binding site of both proteins (Smith *et al.*, 1984; Alshuth *et al.*, 1985).

Table II lists the variation among the different amino acids. Amino acids in the membrane-spanning part can be grouped into three categories. 40–100% conservation is found for Leu, Pro, Thr, Ala, Tyr, Lys, Asp, Trp, Arg and Gly. Conservation between 7 and 30% is found for Ser, Met, Ile and Val, whereas no His, Cys (not present in BR), Glu, Asn and Gln residues are conserved. A striking difference between conservation of the isomeric amino acids Leu (62%) and Ile (18%) or the hydroxy amino acids Thr (50%) and Ser (7%), remains unexplained. In addition, the occurrence of three conserved Pro residues in helices 2, 3 and 6 is noteworthy. Some residues, besides the retinal-binding lysine retained in the BR sequence, can be interpreted in functional terms. Asp 141 (Asp 115 in BR) occurs distant from positively charged amino acids and was shown to react with dicyclocarbodiimide in BR (Renthal *et al.*, 1985). It is a good candidate for the second postulated negative charge in the retinal environment of both proteins (Nakanishi *et al.*, 1980). The counter ion for the Schiff base could be the conserved Asp 238 in HR. Also striking is the conservation of four Trp residues which are the most probable candidates for interaction with retinal in BR (Polland *et al.*, 1986).

Table III. Helical wheel analysis of HR and BR

Helix	Number of amino acid residues	Total score	Hydrophobic segment (score)	Hydrophilic segment (score)
(A) Halorhodopsin				
HR_1	24	46	23–14 (40)	15–22 (6)
HR_2	24	34	2–19 (36)	20–1 (–2)
HR_3	24	22	5–23 (26)	24–4 (–5)
HR_4	24	34	2–19 (36)	20–1 (–2)
HR_5	25	38	13–4 (38)	5–12 (0)
HR_6	24	34	3–21 (0)	22–2 (–6)
HR_7	24	28	19–10 (33)	11–18 (–5)
(B) Bacteriorhodopsin				
BR_1	24	25	9–1 (31)	2–8 (–6)
BR_2	25	30	16–8 (33)	9–15 (–3)
BR_3	25	27	16–8 (30)	9–15 (–3)
BR_4	24	37	8–24 (40)	1–7 (–3)
BR_5	25	35	12–3 (32)	4–11 (3)
BR_6	25	32	16–4 (39)	5–15 (–7)
BR_7	25	43	24–16 (50)	17–23 (–7)

The amino acids of the helices were projected on a circle with 100° spacing between subsequent residues (Schiffer and Edmundson, 1967) resulting in 18 positions on the circle. The total score of segments comprising between 6 and 18 positions were summed using the score value of Kyte and Doolittle and compared with the sum of the residual positions of the helical wheel. In the table the total scores of each helix in HR and BR are given, together with the segmentation of the helical wheel giving maximal difference, i.e. highest amphipathy of each helix.

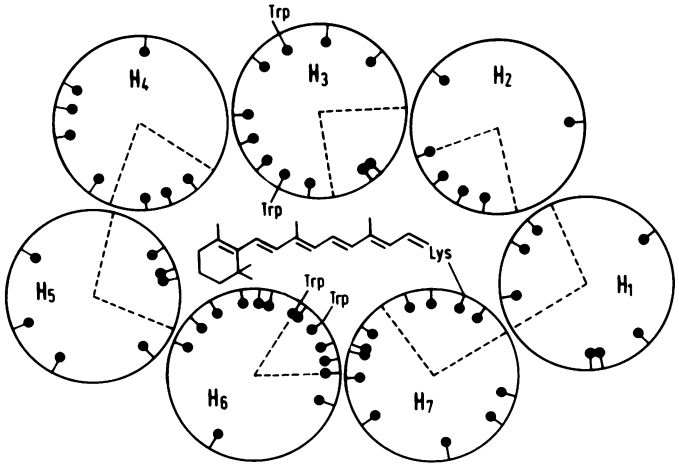

Fig. 7. Conservation of the intrahelical space in the two ion pumps. The seven helical segments of the structural model in Fig. 5 were arranged in the way that the analogous helices are thought to be located in bacteriorhodopsin. The broken lines in helices H_1-H_7 indicate the most hydrophilic segment obtained by helical wheel analysis (Table III) and the dots depict conserved amino acid residues. Alternatively to the presentation in Figure 5 helix 3 was shortened by two amino acids at its beginning and three amino acids at its end. The helical wheel plot then placed all but one conserved amino acid including Trp 106 into the intrahelical hydrophilic space.

Proton conduction in BR has been suggested to involve residues Glu 74, Tyr 79, Tyr 83, Tyr 57, Tyr 150 and Asp 115/212, which can be arranged in a secondary structural model as a hydrogen bridge chain leading half-way through the protein (Merz and Zundel, 1981). Only a few residues corresponding to these positions are conserved in HR, therefore excluding the possibility of forming a hydrogen bridge chain.

Table IV. Comparison of various opsins at a protein and DNA level

Method of comparison	hop	bop	Humops	Bovops	Droops A	Droops B
Relate hop segment 25	–	19.2	0.2	0.5	0.8	0.8
Relate bop segment 25	19.2	–	0.9	0.8	–0.1	0
Align hop penalty 4	–	9.3	0.1	0.9	2.1	3.5
Align hop penalty 19	–	22.6	0.3	0	0.8	–0.2
Align bop penalty 4	9.3	–	0.4	0.4	3.3	2.8
Align bop penalty 19	22.6	–	1.4	1.7	1.2	1.6
% homology of hop DNA	–	51.2	39.4	40.4	38.7	39.5
% homology of bop DNA	51.2	–	40.4	40.8	41.2	40.8

Align and relate programs of protein identification resource, NBRF, Washington, DC with the mutation data matrix from Dayhoff (Dayhoff *et al.*, 1979) and the Gap-program described by Needleman and Wunsch (Needleman and Wunsch, 1970) were used. The similarities of the proteins are expressed in SD units, that of the DNA in percent homology [bacterioopsin -bop- (Dunn *et al.*, 1981); human rhodopsin -Humops- (Nathans and Hogness, 1984); bovine rhodospin -Bovops- (Nathans and Hogness, 1983); *Drosophila* ninaE gene -Droops A- (O'Tousa *et al.*, 1985) and *Drosophila melanogaster* rhodopsin -Droops B- (Zuker *et al.*, 1985)]. It should be noted that proteins can be compared even between very distant members of a gene family provided SD units >5 are found. The statistical homology at the DNA level is 25% but depends on the GC content ($25 + \frac{1}{4}|AT-GC|$).

Discussion

The similarities in the primary structures of HR and BR are suggestive of a similar helix arrangement in both proteins. In the case of BR the ring-like arrangement of the helices 1–7 has been proposed and an angular orientation of the helices postulated, in such a way that an inside-out protein is achieved where the more hydrophobic residues are in contact with the lipid phase and the hydrophilic residues form the intrahelical space (Engelman and Zaccai, 1980). Table III summarizes the result of a helical wheel plot analysis of the amphipathic character of the seven helices in HR and BR. Clearly, the helices in HR as in BR, have an asymmetric angular distribution of hydrophilic amino acids and hydrophilic segments between 60 and 160° are found and shown in Figure 7. The conservation of amino acids between the two proteins is also angle-dependent and the conserved segments coincide, or overlap, with the hydrophilic segments with the exception of helix 3. Forty residues are conserved within the 900° of the inner heptangular space, whereas only 25 residues are conserved in the 1620° of the extrahelical space. It is known that the retinal molecule is embedded inside the BR molecule where it interacts with Trp residues. In fact, three of the four conserved Trp residues are exposed to the intrahelical space (Figure 7).

The surprising similarity of BR and HR structure supports the assumptions that both molecules contain the retinal in a very similar amino acid side chain environment as suggested by Resonance Raman spectra, and that in both systems retinal acts as the light-triggered switch for ion translocation. Only specific binding sites for H^+/OH^- in BR and $Cl^-/Br^-/J^-$ in HR close

to the switch are required, in principle, to explain the different function of the two proteins. The pore surrounded by the helices might then be of different sizes in both proteins and serve the passive diffusion to and from the specific ion binding sites.

Another interesting question concerns the homologies with other retinal proteins, e.g. the visual pigments. Here too, seven transmembrane helices have been suggested for their secondary structure which are connected by loops larger than those observed in either BR or HR (Ovchinnikov, 1982). No charged residues, or amphipathic properties comparable to the ion pumps, were found. The functional parts of these proteins, besides the retinal binding site, indeed must preferentially reside in the helical connection for interaction with transducin, a protein kinase, a phosphatase and the 48-kd protein (Kühn, 1974, 1981; Stryer, 1986). Nevertheless, we searched for homologies of the various retinal proteins at the DNA, as well as at the protein level, by standard programs of sequence comparison (Table IV).

The DNA shows a homology slightly above the randomized sequences of the same GC content. Again, at the protein level, only very minute homologies were found. A test to find helices in the corresponding stretches were found. A test to find helices in the corresponding stretches failed, except for helix 3 in human opsin. It therefore seems that the common features of seven helices of all the opsins are contrasted by the very different function of the external loops in the visual pigments and the ion-conducting transmembrane helices of the pumps. The residual homology required for similar retinal protein interaction (as seen by the red-shifted absorption bands) might involve too few residues in order to be detected by sequence comparison. Thus only the inclusion of the sensor retinal proteins from halobacteria will allow us to prove, or disprove, the idea of common structure principles of retinal proteins beyond the seven helical features.

Materials and methods

Most of the materials and methods are described in the preceding paper (Hegemann *et al.*, 1987). In addition, the following chemicals were purchased. Exonuclease III and nuclease S1 from Boehringer (Mannheim), bovine serum albumin and spermidine from Sigma and plasmids pGem-3 and pGem-4 used for the exonuclease technique as well as the SP6 primer from Promega Biotec (Heidelberg).

The exonuclease technique was used for sequencing the strand of the *hop* gene region with its N-terminal flanking sequences. This was cut out from cosmid pAB H47 with *Asp*718 and *Dde*I and subcloned in the vector system pGem-3 and 4. The plasmids were cut with *Sph*I and *Bam*HI and treated with exonuclease III for 10 varying time periods. Thereafter, treatment with nuclease S1, Klenow polymerase and ligation resulted in a variety of plasmids with shortened inserts. After transformation of *Escherichia coli* DH1 with different DNAs, mini-preparations of DNA from five single clones for each of the 10 time points followed. Plasmids with appropriate insert size were then sequenced with the double-strand sequencing technique (Chen and Seeburg, 1985). Two micrograms of super-coiled plasmid DNA were denatured in 0.2 M NaOH for 5 min at room temperature and then neutralized with 2 M ammonium acetate pH 4.5. After precipitation with two volumes of ethanol the pellet was washed with 1 ml 70% ethanol followed by 1 ml absolute ethanol before drying. The material was then treated as in the dideoxy method of Sanger *et al.* (1977).

Acknowledgements

We would like to thank Dr J.Tittor for preparation of the hydropathy and acrophilicity plots and Dr F.Pfeiffer for help in handling the computer programs for sequence comparison. We further acknowledge the help of V.Gawantka and S.Klostermann in some of the experiments. This work was supported by the Deutsche Forschungsgemeinschaft (Oe 52/16-1).

References

Alshuth,T., Stockburger,M., Hegemann,P. and Oesterhelt,D. (1985) *FEBS Lett.*, **179**, 55–59.

Ariki,M. and Lanyi,J.K. (1984) *J. Biol. Chem.*, **259**, 3504–3510.

Ariki,M., Schobert,B. and Lanyi,J.K. (1986) *Arch. Biochem. Biophys.*, **248**, 532–539.

Betlach,M., Friedman,J., Boyer,H.W. and Pfeifer,F. (1984) *Nucleic Acids Res.*, **12**, 7949–7959.

Chen,E.Y. and Seeburg,P.H. (1985) *DNA*, **4**, 165–170.

DasSarma,S., RajBhandary,U.L. and Khorana,H.G. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 125–129.

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1979) *Atlas of Protein Sequence Structure*, Vol. 5, Suppl. 3, pp. 345–352.

Dennis,P.P. (1985) *J. Mol. Biol.*, **186**, 457–461.

Dunn,R., McCoy,J., Simsek,M., Majumdar,A., Chang,S.H., RajBhandary,U.L. and Khorana,H.G. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 6744–6748.

Engleman,D.M. and Zaccai,G. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 5894–5898.

Engelman,D.M., Steitz,T.A. and Goldman,A. (1986) *Annu. Rev. Biophys. Chem.*, **15**, 321–353.

Hegemann,P., Blanck,A., Vogelsang-Wenke,H., Lottspeich,F. and Oesterhelt,D. (1987) *EMBO J.*, **6**, 259–264.

Henikoff,S. (1984) *Gene*, **28**, 351–359.

Hopp,T.P. (1985) In Alitalo,K., Partanen,P. and Vaheri,A. (eds), *Synthetic Peptides in Biology and Medicine*. Elsevier Science Publishers, Amsterdam.

Jap,B.K. and Kong,S.-H. (1986) *Biochemistry*, **25**, 502–505.

Kühn,H. (1974) *Nature*, **250**, 588–590.

Kühn,H. (1981) *Curr. Top. Membr. Transp.*, **18**, 171–201.

Kyte,J. and Doolittle,R.F. (1982) *J. Mol. Biol.*, **157**, 105–132.

Mankin,A.S., Teterina,N.L., Rubtsov,P.M., Baratova,L.A. and Kagramanova,V.K. (1984) *Nucleic Acids Res.*, **12**, 6537–6546.

May,K. (1986) Diplamarbeit, Univ. München.

Merz,H. and Zundel,G. (1981) *Biochem. Biophys. Res. Commun.*, **101**, 540–546.

Michel,H., Weyer,K.A., Grünberg,H., Dunger,I., Oesterhelt,D. and Lott-Speich,F. (1986) *EMBO J.*, **5**, 1149–1158.

Nakanishi,K., Balogh-Nair,V., Arnaboldi,M., Tsujimoto,K. and Honig,B. (1980) *J. Am. Chem. Soc.*, **102**, 7945–7947.

Nathans,J. and Hogness,D.S. (1983) *Cell*, **34**, 807–814.

Nathans,J. and Hogness,D.S. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 4851–4855.

Needleman,S.B. and Wunsch,C.D. (1970) *J. Mol. Biol.*, **48**, 443–453.

Oesterhelt,D., Hegemann,P., Tavan,P. and Schulten,K. (1986) *Eur. Biophys. J.*, **14**, 123–129.

O'Tousa,J.E., Baehr,W., Martin,P.L., Hirsh,J., Pak,W.L. and Applebury,M.L. (1985) *Cell*, **40**, 839–850.

Ovchinnikov,Yu.A. (1982) *FEBS Lett.*, **148**, 179–191.

Polland,H.J., Franz,M.A., Zinth,W., Kaiser,W. and Oesterhelt,D. (1986) *Biochim. Biophys. Acta*, **851**, 407–415.

Renthal,R., Cothran,M., Espinoza,B., Wall,K.A. and Bernard,M. (1985) *Biochemistry*, **24**, 4275–4279.

Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.

Schegk,E.S., Tittor,J., Lottspeich,F. and Oesterhelt,D. (1986) In Voelter,W., Bayer,E., Ovchinnikov,Y.A. and Ivanov,V.T. (eds), *Chemistry of Peptides and Proteins*. Walter de Gruyter, Berlin, Vol. 3., pp. 259–271.

Schiffer,M. and Edmundson,A.B. (1967) *Biophys. J.*, **7**, 121–135.

Schobert,B., Lanyi,J.K. and Oesterhelt,D. (1986) *J. Biol. Chem.*, **261**, 2690–2696.

Smith,S.O., Marvin,M.J., Bogomolni,R.A. and Mathies,R.A. (1984) *J. Biol. Chem.*, **259**, 12326–12329.

Stryer,L. (1986) *Annu. Rev. Neurosci.*, **9**, 87–119.

Zuker,C.S., Cowman,A.F. and Rubin,G.M. (1985) *Cell*, **40**, 851–858.

Received on 5 November 1986