The EMBO Journal vol.12 no.4 pp.1693-1703, 1993

The probable arrangement of the helices in G protein-coupled receptors

Joyce M.Baldwin

MRC Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 2QH, UK

Communicated by R.Henderson

G protein-coupled receptors form a large family of integral membrane proteins whose amino acid sequences have seven hydrophobic segments containing distinctive sequence patterns. Rhodopsin, a member of the family, is known to have transmembrane alpha-helices. The probable arrangement of the seven helices, in all receptors, was deduced from structural information extracted from a detailed analysis of the sequences. Constraints established include: (1) each helix must be positioned next to its neighbours in the sequence; (2) helices I, IV and V must be most exposed to the lipid surrounding the receptor and helix III least exposed. (1) is established from the lengths of the shortest loops. (2) is determined by considering: (i) sites of the most conserved residues; (ii) other sites where variability is restricted; (iii) sites that accommodate polar residues; (iv) sites of differences in sequence between pairs or within groups of closely related receptors. Most sites in the last category should be in unimportant positions and are most useful in determining the position and extent of lipid-facing surface in each helix. The structural constraints for the receptors are used to allocate particular helices to the peaks in the recently published projection map of rhodopsin and to propose a tentative three-dimensional arrangement of the helices in G protein-coupled receptors.

Key words: G protein-coupled receptor/rhodopsin/sequence comparisons/structure prediction

Introduction

G protein-coupled receptors are integral membrane proteins whose amino acid sequences are characterised by seven hydrophobic segments containing distinctive sequence patterns. Receptors from this family have been found in a wide range of organisms and they are believed to be involved in the transmission of signals across membranes. The receptors bind a signalling molecule on the extracellular side of the membrane and then, following activation, bind and interact with a heterotrimeric guanine nucleotide-binding protein (G protein) on the intracellular side. The activated G protein subsequently initiates a second messenger system of intracellular signalling. Different members of the receptor family respond to quite different ligands, examples being acetylcholine, dopamine, glycoprotein hormones, peptides and other small molecules like adenosine and thromboxane. The binding site for these ligands is in the membrane-embedded part of the protein. Rhodopsin, a visual pigment, is also a member of the family; in this case a light signal induces isomerization of the covalently-bound retinal chromophore. The binding site for the G protein is on the intracellular surface of the protein. In contrast to the great variety of external signalling molecules, there are only a few different types of G protein. Each type of G protein, which is used by many different receptors, initiates a specific type of intracellular signalling; for example an increase in the level of cyclic AMP or a mobilization of intracellular Ca^{2+}. Recent reviews cover various aspects of these receptors, including: visual pigments (Applebury, 1991), rhodopsin–transducin interaction (Hargrave *et al.*, 1993); muscarinic acetylcholine receptors (Hulme *et al.*, 1990); adrenergic receptors (Kobilka, 1992); mutagenesis (Savarese and Fraser, 1992); ligand-binding domain (Oprian, 1992); mechanisms of signalling and desensitization (Dohlman *et al.*, 1991); receptor–effector coupling (Birnbaumer *et al.*, 1990).

All members of the G protein-coupled receptor family are believed to have the same basic structure in the membrane-embedded part of the protein because of their sequence similarities and their common function of interaction with G proteins. It has been shown that in both rhodopsin and the beta-2 adrenergic receptor the N-terminus is on the extracellular side of the membrane and the C-terminus is on the intracellular side (Applebury and Hargrave, 1986; Wang *et al.*, 1989). The transmembrane segments are believed to be alpha-helices, oriented roughly perpendicular to the membrane, as demonstrated for rhodopsin (see Chabre, 1985). It has been widely assumed that the receptors have the same structure as bacteriorhodopsin, an integral membrane protein from *Halobacterium halobium*, whose seven-helical structure is known (Henderson *et al.*, 1990). Although bacteriorhodopsin shares with rhodopsin its response to light through the isomerization of a retinal chromophore, the ground-state conformation of the retinal differs in the two types of protein. Also the response of bacteriorhodopsin is not coupled to G proteins and its sequence shows none of the distinctive patterns of the receptor family. The structures therefore may not be the same.

This paper examines the structural implications of features in the amino acid sequences of G protein-coupled receptors. A probable arrangement for the seven helices in the receptors is based on the results obtained. The prediction is used to allocate particular helices to the peaks in the projection map of rhodopsin that has been obtained by electron crystallography of two-dimensional crystals (Schertler *et al.*, 1993). The combination of the structural information from the sequences and this projection map leads to a tentative structure for the membrane-embedded part of all G protein-coupled receptors.

© Oxford University Press

Table I.

LIGAND	RECEPTOR TYPE: SPECIES		REFERENCES
CATIONIC AMINES			
Acetylcholine	muscarinic m5:man,rat	(98%)	P08912,P08911
	m3:man,rat,pig	(97%)	P20309,P08483,P11483
	m1:man,rat,pig,mouse	(98%)	P11229,P08482,P04761,P12657
	m :drosophila [66% m1]		P16395
	m4:man,rat,chick,xenopus	(96%)	P08173,P08485,P17200,X65865
	m2:man,rat,pig,chick	(98%)	P08172,P10980,P06199,M73217
5-hydroxytryptamine	5HT2:rat,chinese hamster,man	(98%)	P14842,P18599,X57830
	5HT1C:rat,man	(99%)	P08909,M81778
	5HT?:rat [72% 5HT1C rat]		X66842
	5HT1E:man [65% 5HT1D]		M91467
	5HT1D:man(Da),dog,rat [77% 5HT1B]	(91%)	M81589,P11614,M89953
	5HT1B:man(Db),rat,mouse	(97%)	M81590,M89954,Z11597
	5HT1A:man,rat	(98%)	P08908,P19327
	5HT2a,2b:drosophila	(91%)	Z11489,Z11490
	5HT1:drosophila [54% 2b]		P20905
Octopamine	Type-1:drosophila		P22270
Histamine	H2:rat,dog,man	(92%)	P25102,P17124,P25021
	H1:cow		D10197
Dopamine	D1:man,rat	(97%)	P21728,P18901
	D5,D1B:man,rat	(94%)	P21918,P25115
	D2:man,rat,cow,xenopus	(94%)	P14416,P13953,P20288,P24628
	D3:rat,mouse	(99%)	P19020,X67274
	D4:man,rat	(93%)	P21917,M84009
Noradrenaline	alpha-1a:rat,man	(98%)	P23944;P25100
	alpha-1b:rat,hamster,dog	(98%)	P15823,P18841,P11615
	alpha-1c:cow		P18130
	alpha-2a:man,rat,pig	(97%)	P08913,P22909,P18871
	alpha-2b:man,rat	(99%)	P18089,P19328
	alpha-2c:man,rat	(98%)	P18825,P22086
	beta-1:man,rat	(98%)	P08588,P18090
	beta-1:turkey [87% b1 man]		P07700
	beta-2:man,rat,mouse,hamster	(97%)	P07550,P10608,P18762,P04274
	beta-3:rat,mouse	(98%)	P26255,P25962
	beta-3:man [86% b3 rat]		P13945
PEPTIDES			
C5a anaphylatoxin	:man		P21730
Interleukin-8	high affn:man,rabbit	(92%)	P25024,P21109
	low affn:man [89% IL8high]		P25025
fMet-Leu-Phe	R98,R26:man	(99%)	M33537,M33538+P21462
Unknown	FPRL1:man [74% R26,98]		M84562
Unknown	FMPP-R-II:man [100% FPRL1]		P25090
Unknown	FMLP-R-I :man [70% R-II]		P25089
Unknown	RDC1:dog,man	(93%)	P11613,P25106
Somatostatin	SSTR1:man,mouse,rat(100%)		M81829,M81831,X61630
	SSTR2:man,mouse,rat	(97%)	M81830,M81832,M93273
Bradykinin	B2:rat,hum (86%)		P25023,M88714
Angiotensin II	Type-1:rat,cow,man	(97%)	P25095,P25104,Z11162
	Type-1B:rat [99% Type-1 rat]		X64052
Endothelin 1	ET-a:man,cow,rat	(99%)	P25101,P21450,M60786
Endothelin 1,2,3	ET-b:man,rat,cow	(98%)	P24530,P21451,D90456
Neuromedin B	:rat,man	(93%)	P24053,M73482
Gastrin releasing peptide	:mouse,man	(96%)	P21729,M73481
Bombesin-like	PU18:guinea-pig [68% GRP]		X67126
Neurotensin	:rat		P20789
Thyrotropin releasing hormone	:mouse,rat	(98%)	P21761,X64630
Cholecystokinin	CCK-A:rat		M88096
	CCK-B:rat [67% CCKA]		M99418
Gastrin	:dog [93% CCKB]		M87834
Neuropeptide Y	Y1:man,rat	(97%)	P25929,P21555
	Y2:cow [25% Y1]		P25930
	:drosophila [33% Y1,41% NK2]		P25931
Unknown	NKD:drosophila [54% DTKR]		M77168
Unknown	DTKR:drosophila [49% NK3]		X62711
Unknown	C38C10.1:c.elegans [39% NK1,36% DTKR]		Z19153
Unknown	:mouse [39% tachykinins]		M80610
Opioids	:man [93% NK3]		M84605
Neuromedin K(neurokinin B)	NK3:man,rat	(96%)	M73482,P16177
Substance K(neurokinin A)	NK2:man,rat,cow,mouse	(93%)	P21452,P16610,P05363,X62933
Substance P	NK1:man,guinea-pig,rat,mouse	(97%)	P25103,X64323,P14600,X62934
Arginine vasopressin	V2 :man,rat	(94%)	Z11687,Z11932
	V1a:rat		Z11690
Oxytocin	:man		X64878
Thrombin	:hamster,rat	(93%)	M80612,M81642
	:man [86% rat]		P25116
OTHER MOLECULES			
Platelet activating factor	:man,guinea-pig	(91%)	P25105,P21556
Unknown	RTA:rat		P23749
Unknown	mas-oncogene:man,rat	(97%)	P04201,P12526
Prostaglandin E2	EP3:mouse		D10204
Thromboxane A2	:man		P21731
Unknown	edg-1:man		P21453
Cannabinoid	:rat,man	(99%)	P20272,P21554
	:man [57% to above]		(aa)
Adenosine	A1:dog,cow,rat	(97%)	P11616,X63592,P25099
	A2:dog [62% A1 dog]		P11617
	A2b:rat [73% A2 dog]		M91466
	tgpcr1/A3:rat [60% A1]		X59249+M94152
	:man [85% tgpcr1]		(aa)
Unknown	UL33:human cytomegalovirus		P16849
Unknown	US27:human cytomegalovirus		P09703
Unknown	US28:human cytomegalovirus		P09704
GLYCOPROTEIN HORMONES			
Lutropin-choriogonadotropin	:man,rat,pig,mouse	(92%)	P22888,P16235,P16582,M81310
Thyrotropin	:man,rat,dog	(92%)	P16473,P21463,P14763
Follicle stimulating hormone	:man,rat	(95%)	P23945,P20395
LIGHT-INDUCED CHANGE IN BOUND RETINAL			
light;	rod rhodopsin:man,cow,sheep,mouse	(94%)	P08100,P02699,P02700,P15409
	rod rhodopsin:chick [88% cow]		P22328
	rod rhodopsin:goldfish [82% cow]		(cc)
	rod rhodopsin:lamprey [80% cow]		P22671
	rod rhodopsin:sandgoby [82% cow]		(bb)
	green pigment:chick [66% Rh chick]		M92038

green pigments 1,2:goldfish (88%)	(cc)
pigment 467:gecko	M92035
blue cone pigment:man [44% Rh man]	P03999
violet pigment:chick [80% blue man]	M92039
blue pigment:chick [50% blue man]	M92037
blue pigment:goldfish [69% blue chick]	(cc)
red cone pigment:man [43% Rh man]	P04000
red pigment:chick [86% red man]	P22329
red pigment:goldfish,astyanax fasciatus (89%)	(cc) ,P22332
green cone pigment:man [93% red man,44% green chick]	P04001
pigment 521:gecko [78% green man]	M92036
green pigments 1,2:astyanax fasciatus (93%)	P22330,P22331
insect opsin1:calliphora [88% opsin1 dros]	P22269
insect opsin1:drosophila	P06002
insect opsin2:drosophila [70% opsin1]	P08099
insect opsin3:drosophila	P04950
insect opsin4:drosophila [72% opsin3]	P08255
rhodopsin:octopus	P09241
rhodopsin:squid [75% octopus]	P24603
:limulus [55% opsins1,2]	L03781

Author references are given as the accession numbers in the Swissprot (P) and EMBL (M, X, Z, D, L) databanks. Others: (aa) S.Munro, manuscripts in preparation; (bb) Archer *et al.* (1992); (cc) Johnson *et al.* (1993).
A figure in square brackets, where given, is the homology (% identity) over the seven transmembrane segments between the sequence and the named related one. A figure in round brackets is the homology within the pair or group of sequences described on the same line. Pairs or groups of sequences with homologies >90% were used in the analysis of differences between closely related sequences.

Results and Discussion

The database that has been set up for this analysis contains 204 sequences of G protein-coupled receptors. It includes sequences for 76 cationic amine receptors, 32 visual pigments, nine glycoprotein hormone receptors, 66 receptors for peptides and 21 receptors for other small ligands. Some of the sequences are of receptors whose ligands are not yet identified. The receptor proteins in the database are listed in Table I, classified by the type of ligand they bind.

Alignment of the sequences

Hydrophobicity plots of all sequences in the family indicate the presence of seven transmembrane segments, separated by loops that differ in length between different members. The 204 sequences are convincingly aligned within the seven segments from patterns of residues that are characteristic of each of the segments, even though the percentage identity between distantly related members of the family over the seven transmembrane segments is as low as 20%. At least some, and usually all, of the residues in each pattern are present in the corresponding helix for all clear members of the family; a few sequences of receptors for non-peptide, non-cationic amine ligands are not clearly aligned in segment V. The characteristic patterns of residues are shown in Figure 1, which gives their relative positions in the sequence and the percentage of the sequences in which each residue occurs. The percentages have been determined using a set of 105 'unique' sequences; this set includes one example only from each group of sequences with transmembrane segment identity of 85% or more (i.e. one entry for each line in Table I, with the exception of a few lines where the homology to a sequence in another line, given in square brackets, is 85% or more). This unique set comprises sequences for 32 cationic amine receptors, 22 visual pigments, three glycoprotein hormones, 33 peptide receptors and 15 receptors for other small ligands.

The sequences of the olfactory receptors that are now known to be members of the family (Buck and Axel, 1991) are not included in the current analysis. These olfactory receptor sequences can be clearly aligned with those of the main family in all segments except segment VI, where all the usual residues are absent. The sequences of the secretin receptor and other related proteins (see Ishihara *et al.*, 1992), which are known to be coupled to G proteins but have none of the residues that are characteristic of the well known large family, are not considered here.

Choice of helical segments and numbering scheme

As the lengths of the N-terminal regions and of the loops connecting the helices vary between different members of the family, an integrated numbering scheme for positions in the structure is necessary. The following procedure has been used to define the start and finish of each helical segment and to allocate the numbering scheme to the residues within each segment. Different types of residue have been classified as being in one of two classes: (A) a residue that could be in contact with the lipid surrounding the intramembrane part of the protein; (B) a residue that would not be expected to be in contact with lipid. The classification used is compatible with structural features observed in membrane protein structures that are known, for example those of the photosynthetic reaction centre (Deisenhofer and Michel, 1989) and bacteriorhodopsin (Henderson *et al.*, 1990). Class A includes all those residues usually regarded as hydrophobic plus serine, threonine and tyrosine. Serine and threonine residues can satisfy their hydrogen bonding potential by bonding to the main chain of a helix and thus could be on its lipid-facing surface. Tyrosine residues have been observed to be on lipid-facing surfaces in the structure of bacteriorhodopsin (Henderson *et al.*, 1990) and they are also found on lipid-facing surfaces in bacterial porins (Weiss *et al.*, 1991; Cowan *et al.*, 1992), where they mainly occur close to the boundaries between the polar and non-polar parts of the membrane. Residues in class B include all charged residues and all those capable of forming more than one hydrogen bond; the term polar will be used to describe these although serine, threonine and tyrosine have been excluded. The distribution of these two classes of amino acid over the seven segments, as determined using the 105 unique sequences from Table I, is shown in Figure 1. Features of the sequences of each segment are shown in vertical lines that are numbered upwards for segments I, III, V and VII and downwards for segments II, IV and VI; i.e. the intracellular surface of the membrane is at the top of the diagram. The positions in each helical segment have been classified as being: always occupied by a class A amino acid (blank); occupied by a class B amino acid in a few sequences (++); occupied by a class B amino acid in >10% of the sequences (##).

The numbering scheme allows for each helix to include 26 residues. The central position numbers (13 and 14) of each helix were assigned to the middle of the region that has most sites that could be in contact with lipid; the

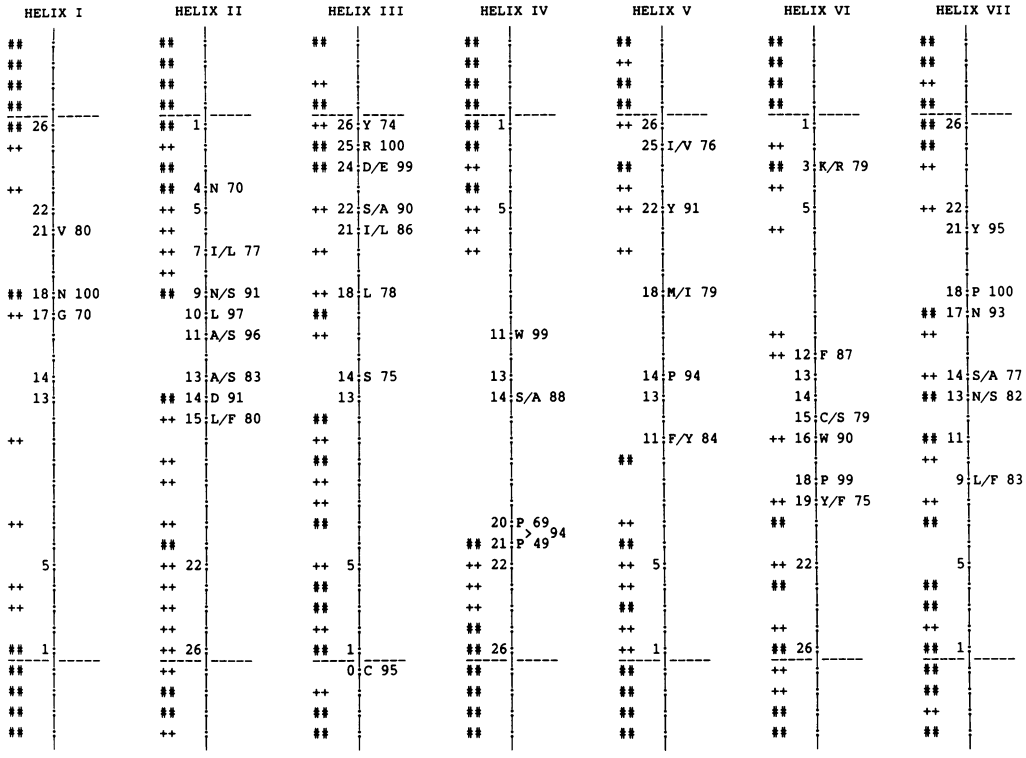

Fig. 1. Residues that are characteristic for each helix and the distribution of polar residues. Positions in each helix are numbered on the left, upwards for segments I, III, V, VII and downwards for segments II, IV and VI; the intracellular surface of the membrane is at the top of the figure. The symbols on the left of the position numbers indicate that a site is occupied by a polar residue in a few sequences (++) or in >10% of the sequences (# #). No symbol on the left indicates that there is never a polar residue at the site. ('Polar' residues include Asp, Asn, Glu, Gln, His, Arg, Lys; all others are regarded as able to face the lipid.) The highly conserved residues in each helix are labelled to the right. The percentage occurrence of the labelled residue (or one of the labelled pair) in 105 unique sequences is shown to the right of the label.

remaining helix positions between 1 and 26 were assigned accordingly. If the helices are approximately perpendicular to the membrane then ~18 residues (5–22) would span the region in which polar residues should not be facing out into the lipid; sites of polar residues would concentrate at either end of the middle region of 18 sites. This is observed for most of the helices, the change in character of the residues at either end of helix III being less obvious. All of the helices except IV have some polar-accommodating site in the middle of the most hydrophobic region; helices II, III and VII have many. The best choice of the segments 1–26 for each helix is uncertain within a few residues, and the relative depths within the membrane of the centre residues of the different helices is uncertain within a turn of the helix. In some of the segments, the number of residues that actually have an alpha-helical conformation may turn out to be <26.

Structural information from comparisons of the sequences

Lengths of inter-helical loops

With the helical regions specified as positions 1–26, the lengths of the connecting loops between the helices and the lengths of the N- and C-terminal regions in the set of receptor sequences vary between the values given in Table II. It can be seen that each type of inter-helical loop is quite short in certain members of the family. As all members of the family are expected to have the same arrangement of the helices, this indicates that each helix is positioned in three dimensions next to those closest to it in the one-dimensional sequence.

Table II. Range of lengths of inter-helical loops over the whole receptor family

Intra/extra	E	I	E	I	E	I	E	I
Region	N-TER	I-II	II-III	III-IV	IV-V	V-VI	VI-VII	C-TER
Min size	6	5	13	10	12	12	8	12
Max size	394	11	22	18	43	420	20	162

The shortest receptor has 319 amino acids and the longest 834.

Structural features of each helix

The features of the helices are best summarized by plotting them around helical wheels. Figure 2 shows such wheels for the middle 18 positions of the seven helices. The alpha-carbon positions are marked by squares whose size indicates depth in the structure, with the larger squares uppermost (viewed from the intracellular surface). The positions have been given an amino acid label if more than half the 105 unique sequences have the same residue there or if the site is interesting for other reasons. In the latter case, the position is labelled with the most common amino acid.

(i) *Positions of the most conserved residues.* The positions of the characteristic residues by which the sequences are aligned are marked by spokes on the wheels in Figure 2. Solid spokes mark sites that have the same amino acid or one of a very closely related pair of amino acids, in 85% of the 105 unique sequences; open spokes mark sites at which this occurs in 75% of the sequences. Single spokes correspond to the same residue occurring in 66% of the sequences or one of a less closely related pair occurring in 75%. These characteristic residues are expected to play a

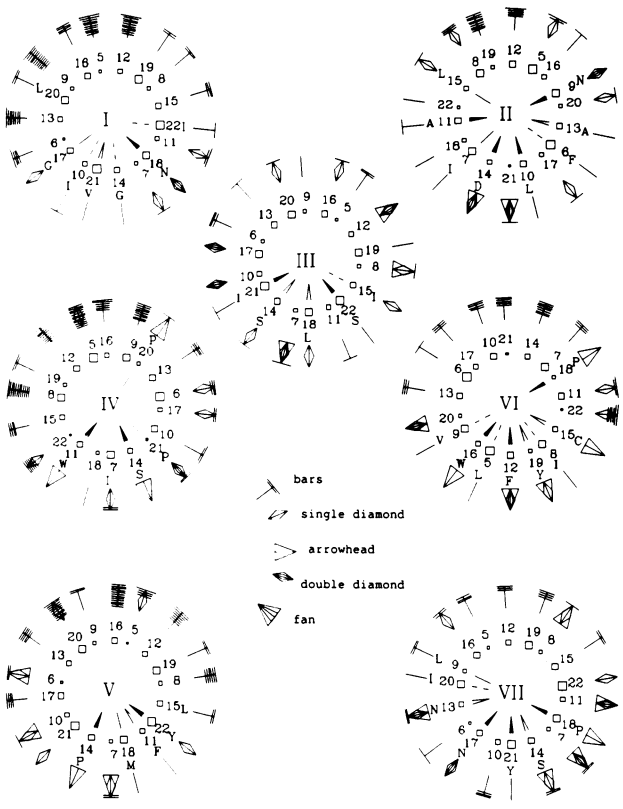

Fig. 2. Sequence comparison information summarized around helical wheels for each of the seven helices. Squares mark alpha-carbon positions for the middle 18 residues of each helix; size indicates depth in the structure, viewed from the intracellular surface. Positions are numbered as in Figure 1 and are given an amino acid label if more than half the sequences have that particular residue. Inner spokes mark positions of characteristic residues as classified in section (i) of the text. Broken inner spokes mark other sites where variation is limited as described in section (ii) of the text. Diamonds on outer spokes mark sites where polar residues are found; single diamonds correspond to ++ in Figure 1, double diamonds to # #. (See legend to Figure 1 and section (iii) of the text.) Bars on the outer spokes mark the positions where differences occur between very closely related sequences as described in section (iv) of the text. Each helix has been oriented so that its lipid-facing surface is at the top of the figure, after combining the evidence discussed in sections (i)–(iv). Arrowheads on the outer spokes mark the positions of sites involved in interactions with ligands; see data listed in Table III.

role in maintaining the structure of the molecule so that it can bind the G protein, this being the function that all the receptors have in common. Thus these residues will probably face either towards the inside of the molecule or towards another helix. Conserved prolines, which will cause kinks in the helices, may be in any position in the structure. Each of the seven helices has its characteristic residues other than prolines concentrated on one side of the wheel. The majority of the highly conserved residues are in the intracellular half of the protein.

(ii) *Positions where variability is restricted.* There are very few sites other than those of the characteristic residues that have any restriction on the size of the residue that they can accommodate or on the number of different types of amino acids that are seen there. The additional sites where variation is limited are marked in Figure 2 by a broken inner spoke on the wheel and are labelled with the most common amino acid. Sites are designated as allowing only limited variation if they satisfy one or both of two criteria: first, if less than nine different amino acids are found there; secondly, if a restricted range of side-chain sizes is found there. Residue sizes are categorized as being either small, medium or large. Sites satisfy the second criterion if they are observed to accommodate residues from only one size category, or from two neighbouring categories. These slightly restricted sites would be expected to face the centre of the molecule or be at interfaces between helices; they are observed to be on the same faces of the helices as the characteristic residues.

(iii) *Positions that accommodate polar residues.* Diamonds on the outside of the wheels in Figure 2 show the sites at which polar residues are seen. Single diamonds mark sites at which a few of the sequences have polar residues [classification (++) of Figure 1] and double diamonds mark sites where >10% of the sequences have polar residues (classification (# #) of Figure 1). The face of each helix that could be in contact with the surrounding lipid is that which has no polar sites. Positions numbered 5, 6, 7 or 22, 21, 20 (i.e. quite near the membrane surfaces) could face the lipid and still accommodate polar residues that reach to the hydrophilic surface. Also the choice of the centre residues of each segment is defined only to within a few residues. It is, however, clear that for most of the helices one face never has a polar residue in the middle of the membrane. Helices I, IV, V and VI each have a large surface with no polar residues; helices II, III and VII have smaller ones.

(iv) *Positions of differences between closely related sequences.* The sequences listed on the same line in Table I are from pairs or groups of receptors with essentially the same function, but from different species. The percentage identities within these pairs or groups, given in the table in round brackets, are extremely high. There are 49 different groups with homologies of 90% or more, made up from a total of 129 sequences. If the pairs and groups of receptors have essentially the same function as each other, most of the sequence differences between them must occur at unimportant sites; in the transmembrane segments the differences must be on the outside of the molecule, facing the lipid. A few of the sequence differences may correspond to subtle differences in function, in which case the sites of these would be facing inside the molecule. The positions of differences between very closely related sequences are indicated in Figure 2. For each available pair or group, the sites where the few differences occur have been recorded by marking a bar on the outer spoke at the site; each pair or group contributes between one and 12 bars to the diagram, depending on its degree of homology. The shorter bars indicate conservative sequence differences; the longer bars indicate more substantial differences. In this way a mapping of the probable lipid-facing surface of each helix has been built up. Helices I, II, IV, V and VI show clear concentrations of differences on one side of the helix; helix III has fewer such differences and in helices III and VII the differences are less segregated on one side. If the helices are put in order of decreasing lipid-facing surface area, the order is I/IV/V followed by VI, then II/VII, with III appearing to have very little lipid-facing area and then only at the extracellular end of the helix.

The analysis of the sequence data in terms of close-pair differences, polar-accommodating sites and sites of restricted overall variation has shown that the inside surfaces of the helices can be discriminated from the lipid-facing surfaces. This is demonstrated in Figure 2 where each helix has been

Table III.

Hlx:n	Residue			Reference
Sites of functional importance in visual pigments				
III:0	C	Involved in disulphide bridge		Karnik and Khorana (1990)
III:3	E	Counterion to Schiff base, vertebrates		Oprian (1992) (review)
III:12	E	Mutation Q,D blue-shifts chromophore		Nakayama and Khorana (1991)
IV:14	A/S	cause of green/red absorption shift		Neitz *et al.* (1991)
VI:12	F/Y	cause of green/red absorption shift		Neitz *et al.* (1991)
VI:16	W	Mutation Y,F blue-shifts chromophore		Nakayama and Khorana (1991)
VI:20	A/T	cause of green/red absorption shift		Neitz *et al.* (1991)
VII:11	K	Retinal site; protonated Schiff base		Oprian (1992) (review)
Sites identified by labelling experiments				
III:7	D105	m1:rat	binds affinity reagent	Kurtenbach *et al.* (1990)
VII:8	W330	beta:turkey	photochemically labelled	Wong *et al.* (1988)
Sites of mutations that affect agonist and/or antagonist binding				
Hlx:n	Mutation			Reference
II:14	D79A	beta-2:hamster		Strader *et al.* (1987, 1988)
	D79N	beta-2:man		Chung *et al.* (1988)
II:21	A120A	m3:rat		Wess *et al.* (1991)
III:0	C106V	beta-2:hamster		Dixon *et al.* (1987)
III:7	D113N	beta-2:hamster		Strader *et al.* (1987)
	D105N	m1:rat		Fraser *et al.* (1989)
III:8	T148F	m3:rat		Wess *et al.* (1991)
IV:11	W192F	m3:rat		Wess *et al.* (1993)
IV:20	P201A	m3:rat		Wess *et al.* (1993)
V:3	T321A	m3:rat		Wess *et al.* (1991)
V:6	T234A	m3:rat		Wess *et al.* (1991)
	S198A	D1:man		Pollock *et al.* (1992)
V:7	S204A	beta-2:man		Strader *et al.* (1989)
	S199A	D1:man		Pollock *et al.* (1992)
V:10	S207A	beta-2:man		Strader *et al.* (1989)
	S204A	alpha-2a:man		Wang *et al.* (1991)
	S242A	5HT2:man		Kao *et al.* (1992)
	S202A	D1:man		Pollock *et al.* (1992)
V:14	P242A	m3:rat		Wess *et al.* (1993)
VI:16	W503F	m3:rat		Wess *et al.* (1993)
VI:18	P505A	m3:rat		Wess *et al.* (1993)
VI:19	Y506F	m3:rat		Wess *et al.* (1991)
VII:7	Y529F	m3:rat		Wess *et al.* (1991)
	F412N	alpha-2a:man		Suryanarayana *et al.* (1991)
	N385V	5HT1A:man		Guan *et al.* (1992)
	T355N	5HT1B:man		Oksenberg *et al.* (1992)
VII:8	W530F	m3:rat		Wess *et al.* (1993)
VII:11	Y533F	m3:rat		Wess *et al.* (1991)
VII:13	N318K	beta-2:hamster		Strader *et al.* (1987)
VII:14	S319A	beta-2:man		Strader *et al.* (1989)
VII:18	P540A	m3:rat		Wess *et al.* (1993)

oriented so that its lipid-facing surface is toward the top of the figure and its inward facing surface is toward the bottom. This clear discrimination confirms that the segments are indeed helices.

The identification of the inside surfaces of the helices is corroborated by current knowledge of sites that are involved in binding agonists and/or antagonists; such sites are expected to be inside the molecule participating in a ligand binding pocket. The binding pockets for agonists and antagonists may not be identical but they probably overlap. In the case of the visual pigments, some residues are known to affect the absorption spectrum of the retinal chromophore. Table III lists some of the data from site-directed mutagenesis, chemical labelling or other studies that have established the involvement of particular residues. The sites of these residues are marked in Figure 2 by arrowheads; it can be seen that almost all these sites are on the inward pointing faces of the helices. With the exception of helix I, all helices have been shown to contribute to the ligand binding pocket. Almost all the sites involved are in the extracellular half of the transmembrane helices.

The arrangement of the helices

The analysis has also shown that some of the seven helices have distinctive environments. As demonstrated in Figure 2, helices I, IV and V have environments most exposed to

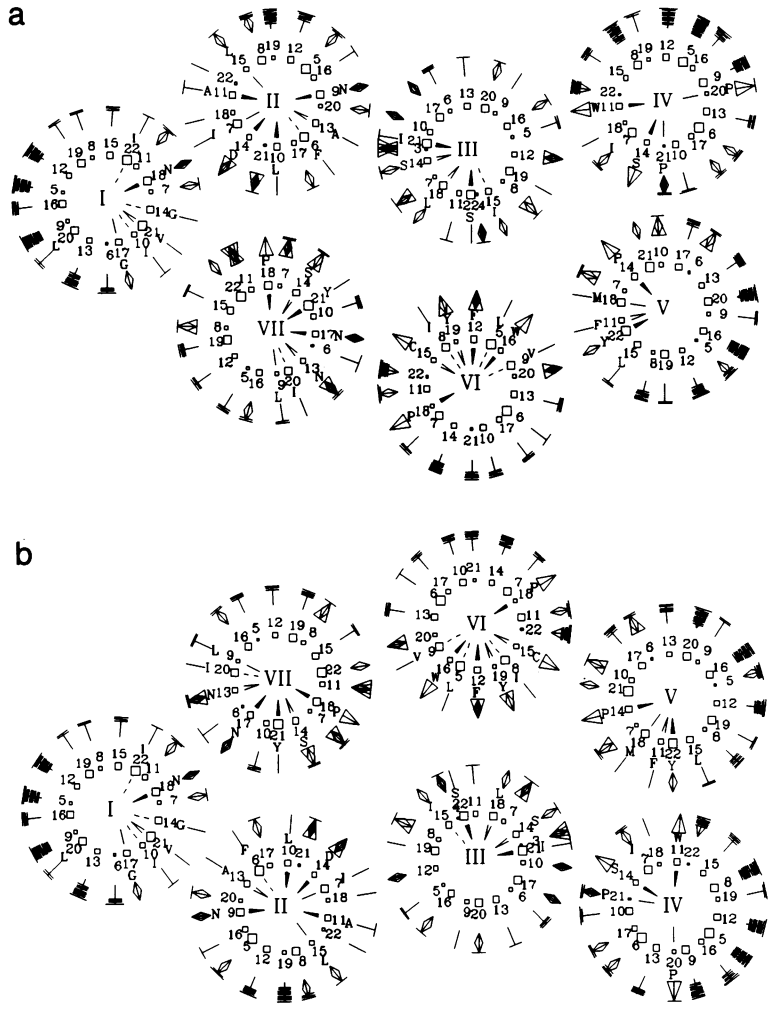

Fig. 3. Possible arrangement of the helices, based on the information summarized in Figure 2 together with the constraint of short inter-helical loops. In (a), the connectivity is clockwise viewed from the intracellular surface; in (b) it is anticlockwise. In both (a) and (b) the individual helices are orientated with lipid-facing surfaces outwards. (a) is the most likely arrangement because of the better disposition of positions III:3 and VII:11 (marked with additional fan symbols; see Figure 2 for key to symbols).

the lipid surrounding the structure; helix III has very little exposure to lipid, particularly at the intracellular end, whilst helices II, VI and VII have intermediate degrees of exposure. Thus the arrangement of the helices in the structure of a G protein-coupled receptor must be such as to put helix III in a position that is well protected from the lipid, whilst leaving helices I, IV and V most exposed to the lipid. The two types of arrangement compatible with these conclusions are illustrated in Figure 3a and b: they differ only in having either clockwise or anticlockwise connectivity as viewed from the intracellular side of the membrane. They are the only types of arrangement that satisfy the following criteria: (i) each helix must be next in space to the two next to it in sequence; (ii) helices I, IV and V must be most exposed to the lipid, helix III must be most buried.

In both Figures 3a and b, each helix has been oriented so that its lipid-facing surface is on the outside of the molecule. In both figures, the characteristic residues of each helix, with the exception of prolines IV:20 and VI:18, are pointing towards other helices. Also, most of the polar residues and the site-directed mutagenesis sites are pointing inwards. The arrangement in Figure 3a can be selected as the most likely one, by considering the established role of certain residues in rhodopsin. The retinal chromophore in rhodopsin is covalently attached to Lys296 (see Oprian, 1992), which is at position VII:11. The positively charged Schiff base at the end of the lysine side chain requires a negative counterion, which has been identified as Glu113, at position III:3 (Zhukovsky and Oprian, 1989; Sakmar *et al.*, 1989; Nathan, 1990). In Figure 3a these positions are suitably disposed to allow such a charge interaction and the retinal chromophore would have space to extend to the right between the helices. The chromophore is known to lie approximately parallel to the membrane (Thomas and Stryer, 1982). In Figure 3b the separation of the two positions is too great and the space for the chromophore would be too restricted.

Evidence that the arrangement in the receptors differs from that of bacteriorhodopsin

This proposed arrangement of the helix axes for the G protein-coupled receptors (Figure 3a) is similar to that seen in the middle of the membrane-embedded part of bacteriorhodopsin (Henderson *et al.*, 1990), although it has been deduced without reference to it. Figure 4a, b and c show the helical wheels placed at the observed positions of the helix axes in bacteriorhodopsin in the intracellular half, the middle and the extracellular half of the structure respectively.

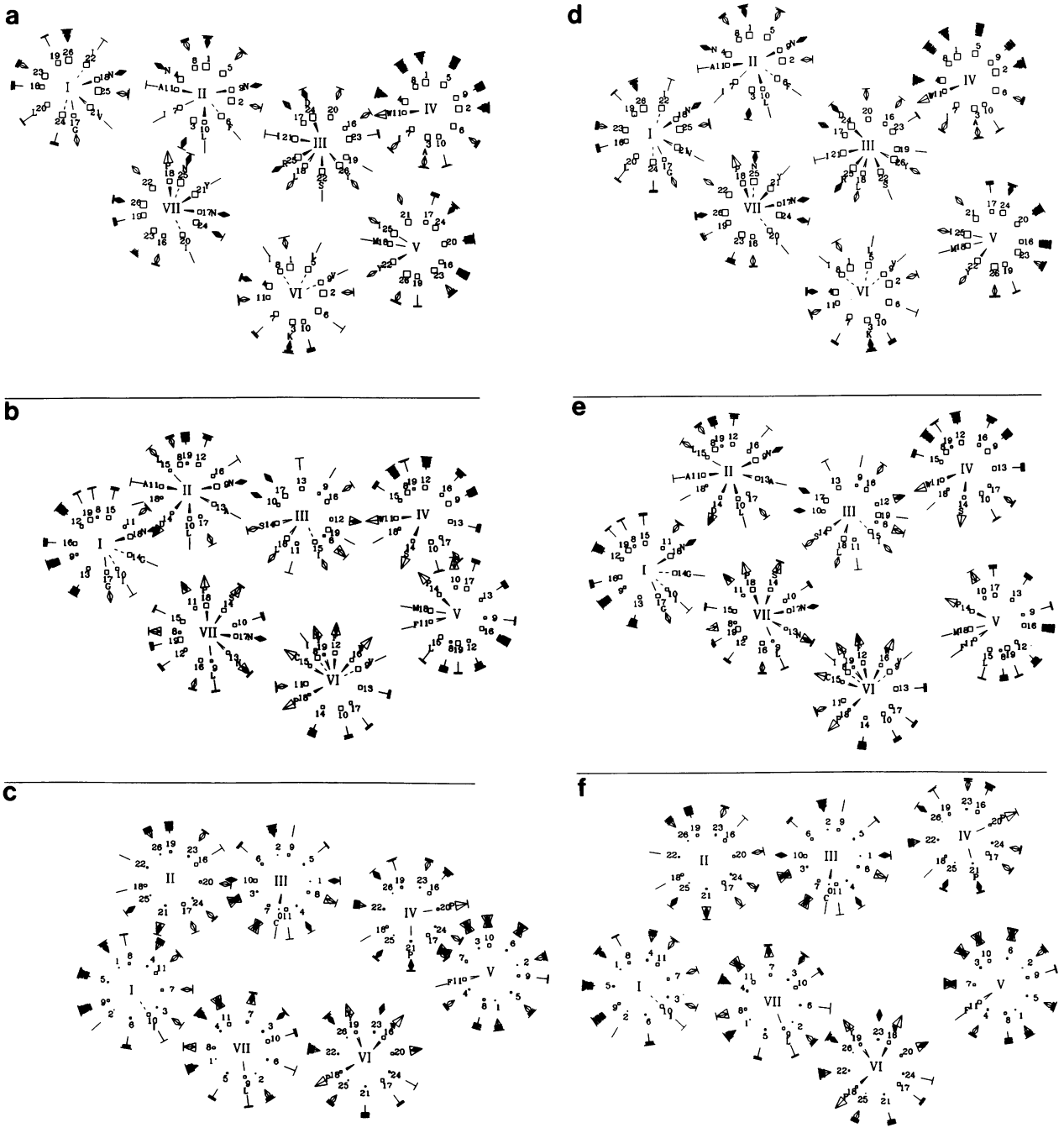

Fig. 4. On the left, the receptor helices positioned as they would be in bacteriorhodopsin (bR). (**a**) 11 residues from the intracellular end of each 26-residue helix, with helix axes positioned as in the intracellular half of bR (14A from middle); (**b**) the middle 12 residues, with axes positioned as the middle of bR; (**c**) 11 residues from the extracellular end of each helix, with axes positioned as the extracellular half of bR (14A from middle). The helices are oriented with lipid-facing surfaces on the outside. In panel c, helix III is separated from helix V by helix IV and residues at positions V:3, V:6, V:7 and V:10 are not suitably placed for interacting wth a small cationic amine ligand bound at position III:7. On the right, the receptor helices positioned as proposed for G protein-coupled receptors. (**d**) 11 residues from the intracellular end of each 26-residue helix; (**e**) 12 residues from the middle of each helix; (**f**) 11 residues from the extracellular ends of each helix. In panel f residues at positions V:3, V:6, V:7 and V:10 are suitably placed for interacting with a small cationic amine ligand bound at position III:7.

There are, however, indications that the positions of the axes in the receptor structure may not change with depth in the same way. The first piece of evidence concerns the disulphide bridge that is present on the extracellular side of the receptors, between the highly conserved cysteine residue at site III.0 and another cysteine residue, which is highly conserved in occurrence, but not in position, in the IV–V loop. The importance of this bridge has been established for rhodopsin (Karnik and Khorana, 1990), muscarinic receptors (Kurtenbach *et al.*, 1990) and for adrenergic receptors (Dohlman *et al.*, 1990), and it is believed to be present in the majority of receptors. The length of the IV–V loop in some members of the family is as short as 12 residues (see Table II) and the number of amino acids between the cysteine in the loop and the end of helix V can be very small; five in all 5-HT2 and dopamine-D2 receptors. Thus the extracellular end of helix V has to be quite near the extracellular end of helix III, which is not the case in the bacteriorhodopsin

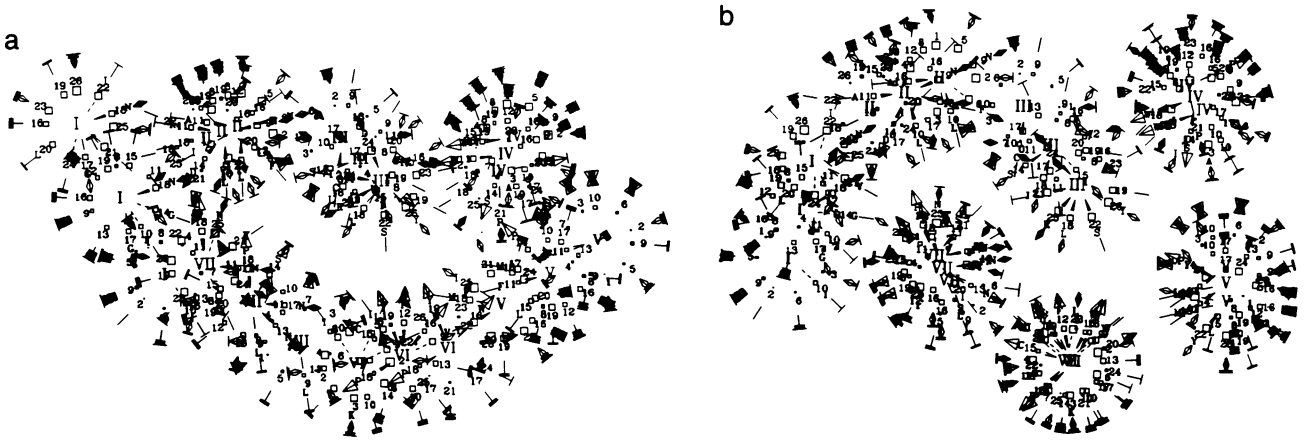

Fig. 5. (**a**) A superposition of Figure 4a, b and c to show the appearance in projection of the bacteriorhodopsin arrangement of helices, viewed from the intracellular surface. (**b**) A superposition of Figure 4d, e and f to show the appearance in projection of the proposed structure of a receptor, viewed from the intracellular surface. This superposition matches the projection structure of rhodopsin (Schertler *et al.*, 1993), which was used to guide the choice of helix axes positions in Figure 4d, e and f as explained in the text.

structure (Figure 4c) where helix IV lies between them. The second piece of evidence concerns the residues implicated in the binding of cationic amines. The amine groups of cationic amine ligands bind to the aspartic acid residue at III:7 (Kurtenbach *et al.*, 1990). Ligands related to adrenaline interact with serine residues at V:7 and V:10 and acetylcholine interacts with threonine residues at V:3 and V:6 (see data in Table III). In Figure 4c, where the helices have been orientated with lipid-facing surfaces on the outside, the relative positions of these residues in helices III and V do not appear appropriate for the binding of these short ligands.

The change in helix axes positions with depth in the structure of bacteriorhodopsin is the consequence of the particular slopes of helices V, VI, VII and I relative to the membrane, helices II, III and IV being approximately perpendicular to the membrane. It seems probable that in the receptors, the helices slope differently relative to the membrane and relative to each other, resulting in an arrangement of helices III, IV and V on the extracellular side of the structure that is more like that in Figure 3a, allowing helices III and V to face each other without helix IV coming between them. The fact that the helix packing is somewhat different in the two types of structure could be associated with the different positions of conserved proline residues. In the bacteriorhodopsin family of seven-helix membrane proteins there are conserved prolines in helices II, III and VI; in the receptor family there are conserved prolines in helices IV, V, VI and VII. The proline in helix VI is on the opposite side of the helix in the two types of structure.

Tentative three-dimensional structure of G protein-coupled receptors

A projection map of rhodopsin at 9 Å resolution has recently been determined by electron crystallography of two-dimensional crystals (Schertler *et al.*, 1993). The structure of rhodopsin seen in this map is clearly different from the structure of bacteriorhodopsin. The suggestion that the slopes of the helices relative to the membrane are different in G protein-coupled receptors from those in bacteriorhodopsin is compatible with the rhodopsin map. Schertler and colleagues have interpreted four resolved peaks in their projection structure to be the density from four helices that are approximately perpendicular to the membrane, and an arc-shaped feature to be the density from three tilted helices that overlap in projection. The conclusions from the sequence comparisons described above can be used to allocate particular helices to the peaks in the projection structure. The proposed arrangement of the helix axes in Figure 3a can then be adjusted at three levels in the membrane to match the projection structure with a three-dimensional interpretation.

The projection map of rhodopsin contains two symmetry related views of the molecule, either of which might be interpreted as the view from the intracellular surface of the membrane. Both interpretations were investigated, but only one gave an arrangement that fitted the established constraints. Figure 4d, e and f show proposed positions of the helix axes in the receptor structure, in the intracellular half, the middle and the extracellular half of the membrane respectively. In Figure 5b the three sets of positions are overlapped to show how they match the projection structure described by Schertler and colleagues, producing four separated peaks that arise from four helices that are nearly perpendicular to the membrane and one curved, elongated peak that arises from the superposition of three helices. (This superposition differs from that of the bacteriorhodopsin arrangement of the helices in Figure 5a.) The allocation of the helices is derived from the arrangement in Figure 3a. The additional criterion that helix III must be more buried at the intracellular end than at the extracellular end has now been included. This condition is satisfied if the most buried end of the curved, elongated peak in the observed projection structure corresponds to the intracellular end of helix III and if the helix slopes away from the centre of the molecule towards the extracellular surface. The extracellular end of helix III must be overlapped in the projection by the intracellular end of helix II and the extracellular end of helix II by the intracellular end of helix I. The other four helices are nearly perpendicular to the membrane so their slopes are less important in specifying the general nature of the structure. The interhelical angles and the distances between the helices that correspond to the axes positions in Figure 4d, e and f are compatible with known structural principles of helix–helix packing (see Chothia, 1984). All the helices pack against their neighbours at small positive angles except

for the packing of helix III against helix IV, which has a small negative angle. The corresponding interaction in bacteriorhodopsin also has a negative angle and its magnitude is similar. The helices in Figure 4d, e and f have been oriented with their lipid-facing surfaces on the outside of the molecule.

The suggested arrangement gives a closely packed structure at the intracellular surface (Figure 4d), where the G protein interaction occurs. The regions of the structure implicated in binding the G protein (see Hargrave, 1991) are close together; these are the highly conserved Asp/Glu-Arg-Tyr (D/E-R-Y) sequence at the end of helix III, the N- and C-terminal portions of the loop between helices V and VI, and the portion of the C-terminus near to helix VII. Presumably a conformational change produced in this region after the binding of ligand enables binding and activation of the G protein.

There is a more open structure in the extracellular half of the protein (Figure 4f), producing a ligand binding pocket. The positions VII:11 and III:3, the sites of Lys296 and Glu113 in rhodopsin, are pointing towards each other, separated by a suitable distance. The binding pocket for the retinal chromophore is a different shape from that in bacteriorhodopsin, which is compatible with the retinal having the 11-*cis* conformation (bent) in the ground state of rhodopsin, whereas it has the all-*trans* conformation (fully extended) in the ground state of bacteriorhodopsin. Ligands such as acetylcholine and noradrenaline are smaller than retinal and would occupy the part of the pocket that extends to the right from the aspartic acid at position III:7 towards helix V where the threonine residues V:3 and V:6 and the serine residues V:7 and V:10 implicated in ligand binding are suitably situated. The other sites involved in ligand binding (Table III) also fall in reasonable positions. The highly conserved cysteine residue at position III:0 is at a suitable distance from the end of helix V.

Conclusions

Structural information has been extracted from the sequences of this large family of seven-helix membrane proteins and different environments for the helices have been detected. The arrangement of the helices in the structure has been deduced by allocating each helix to a position appropriate to the extent of its lipid-facing surface area. Previous analysis of sequence variability between members of this receptor family has produced orientation vectors for the helices (Donnelly *et al.*, 1989) and the helices have also been orientated on the basis of hydrophilic residues (e.g. Hibert *et al.*, 1991). The combination of information used here, in particular the differences between the sequences of very closely related receptors, has proved powerful in emphasizing the differences between the helices, as well as establishing their orientation. Consideration of the proposed arrangement in conjunction with data available from site-directed mutagenesis and other studies suggests that the structure of the receptors is not exactly the same as that of bacteriorhodopsin, although the general arrangement is the same.

The projection map of rhodopsin, obtained experimentally by Schertler *et al.* (1993) also suggests that the structure of the receptors differs from the structure of bacteriorhodopsin. Structural information from the sequences enables the peaks in the map to be allocated to particular helices and suggests a plausible extrapolation to the three-dimensional arrangement in the receptors. This proposed arrangement is still speculative. Three-dimensional crystallographic data is required to determine the structure experimentally. The structural information described here will be more powerful when used in conjunction with a three-dimensional map, even if obtained only at relatively low resolution.

Materials and methods

The 204 sequences summarized in Table I were aligned by eye and incorporated in a database using the homologous sequence editor HOMED, written by Peter Stockwell. For convenience the database was built up in four parts comprising: (i) cationic amine receptors, (ii) visual pigments, (iii) glycoprotein hormone receptors, (iv) peptide receptors, receptors for other small ligands and some with ligands still unknown or unclassified.

Statistical analysis of the sequences in the database was carried out with a specially written program that used the four disk outputs from the HOMED runs as input. This program can analyse all or a specified subset of the sequences, selecting them from one or more of the databases. The start and finish positions of seven segments in the overall aligned dataset can be specified.

(i) For each position in each helix the following information can be obtained for the set of sequences selected: the number of times each type of amino acid occurs; the number of times polar amino acids occur; which receptors have which amino acid; the size range of the amino acids that occur.

(ii) A chart of the percentage identity between all pairs of sequences can be produced. If desired a score matrix can be used instead of a 0/1 identity scheme.

(iii) The program can also give a list of the positions at which there is a difference in sequence between a specified pair of receptors. It will list positions and name the two different amino acids, for any number of specified pairs. It can also be used to list the positions at which all receptors of a specified group have a conserved amino acid.

Acknowledgements

I thank my colleagues Richard Henderson and Gebhard Schertler for helpful discussions, Peter Stockwell of Otago University, New Zealand for making his Homologous Sequence Editor available and authors for making their sequences available through the databanks.

References

Applebury,M.L. (1991) *Curr. Opin. Neurobiol.*, **1**, 263–269.
Applebury,M.L. and Hargrave,P.A. (1986) *Vision Res.*, **26**, 1881–1895.
Archer,S.N., Lythgoe,J.N. and Hall,L. (1992) *Proc. R. Soc. Lond. B*, **248**, 19–25.
Birnbaumer,L., Abramowitz,J. and Brown,A.M. (1990) *Biochem. Biophys. Acta*, **1031**, 163–224.
Buck,L. and Axel,R. (1991) *Cell*, **65**, 175–187.
Cowan,S.W., Schirmer,T., Rummel,G., Steiert,M., Ghosh,R., Pauptit, R.A., Jansonius,J.N. and Rosenbusch,J.P. (1992) *Nature*, **358**, 727–733.
Chabre,M. (1985) *Annu. Rev. Biophys. Chem.*, **14**, 331–360.
Chothia,C. (1984) *Annu. Rev. Biochem.*, **53**, 537–572.
Chung,F.-Z., Wang,C.-D., Potter,P.C., Venter,J.C. and Fraser,C.M. (1988) *J. Biol. Chem.*, **263**, 4052–4055.
Deisenhofer,J. and Michel,H. (1989) *Science*, **245**, 1463–1473.
Dixon,R.A.F., Sigal,I.S., Candelore,M.R., Register,R.B., Scattergood,W., Rands,E. and Strader,C.D. (1987) *EMBO J.*, **6**, 3269–3275.
Dohlman,H.G., Caron,M.G., DeBlasi,A., Frielle,T. and Lefkowitz,R.J. (1990) *Biochemistry*, **29**, 2335–2342.
Dohlman,H.G., Thorner,J., Caron,M.G. and Lefkowitz,R.J. (1991) *Annu. Rev. Biochem.*, **60**, 653–688.
Donnelly,D., Johnson,M.S., Blundell,T.L. and Saunders,J. (1989) *FEBS Lett.*, **251**, 109–116.
Fraser,C.M., Wang,C.-D., Robinson,D.A., Gocayne,J.D. and Venter,J.C. (1989) *Mol. Pharmacol.*, **36**, 840–847.
Guan,X.M., Peroutka,S.J. and Kobilka,B.K. (1992) *Mol. Pharmacol.*, **41**, 695–698.
Hargrave,P.A. (1991) *Curr. Opin. Struct. Biol.*, **1**, 575–581.
Hargrave,P.A., Hamm,H.E. and Hofmann,K.P. (1993) *BioEssays*, **15**, 1–8.

Henderson,R., Baldwin,J.M., Ceska,T.A., Zemlin,F., Beckmann,E. and Downing,K.H. (1990) *J. Mol. Biol.*, **213**, 899–929.
Hibert,M.F., Trump-Kallimeyer,S., Bruinvels,A. and Hoflack,J. (1991) *Mol. Pharmacol.*, **40**, 8–15.
Hulme,E.C., Birdsall,N.J.M. and Buckley,N.J. (1990) *Annu. Rev. Pharmacol. Toxicol.*, **30**, 633–673.
Ishihara,T., Shigemoto,R., Mori,K., Takahashi,K. and Nagata,S. (1992) *Neuron*, **8**, 811–819.
Johnson,R.L., Grant,K.B., Zankel,T.C., Boehm,M.F., Merbs,S.L., Nathans,J. and Nakanishi,K. (1993) *Biochemistry*, **32**, 208–214.
Kao,H.-T., Adham,N., Olsen,M.A., Weinshank,R.L., Branchek,T.A. and Hartig,P.R. (1992) *FEBS Lett.*, **307**, 324–328.
Karnik,S. and Khorana,H.G. (1990) *J. Biol. Chem.*, **265**, 17520–17524.
Kobilka,B. (1992) *Annu. Rev. Neurosci.*, **15**, 87–114.
Kurtenbach,E., Curtis,C.A.M., Pedder,E.K., Aitken,A., Harris,A.C.M. and Hulme,E.C. (1990) *J. Biol. Chem.*, **265**, 13702–13708.
Nakayama,T.A. and Khorana,H.G. (1991) *J. Biol. Chem.*, **266**, 4269–4275.
Nathan,S.J. (1990) *Biochemistry*, **29**, 9746–9752.
Neitz,M., Neitz,J. and Jacobs,G.H. (1991) *Science*, **252**, 971–974.
Oprian,D.D. (1992) *J. Bioenerg. Biomembr.*, **24**, 211–217.
Oksenberg,D., Marsters,S.A., O'Dowd,B.F., Jin,H., Havlik,S., Peroutka,S.J. and Ashkenazi,A. (1992) *Nature*, **360**, 161–163.
Pollock,N.J., Manelli,A.M., Hutchins,C.W., Steffey,M.E., MacKenzie,R.G. and Frail,D.E. (1992) *J. Biol. Chem.*, **267**, 17780–17786.
Sakmar,T.P., Franke,R.R. and Khorana,H.G. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 8309–8313.
Savarese,T.M. and Fraser,C.M. (1992) *Biochem. J.*, **283**, 1–19.
Schertler,G.F.X., Villa,C. and Henderson,R. (1993) *Nature*, in press.
Strader,C.D., Sigal,I.S., Register,R.B., Candelore,M.R., Rands,E. and Dixon,R.A.F. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 4384–4388.
Strader,C.D., Sigal,I.S., Candelore,M.R., Rands,E., Hill,W.S. and Dixon,R.A.F. (1988) *J. Biol. Chem.*, **263**, 10267–10271.
Strader,C.D., Candelore,M.R., Hill,W.S., Sigal,I.S. and Dixon,R.A.F. (1989) *J. Biol. Chem.*, **264**, 13572–13578.
Suryanarayana,S., Daunt,.DA., von Zastrow,M. and Kobilka,B.K. (1991) *J. Biol. Chem.*, **266**, 15488–15492.
Thomas,D.D. and Stryer,L. (1982) *J. Mol. Biol.*, **154**, 145–157.
Wang,H.Y., Lipfert,L., Malbon,C.C. and Bahouth,S. (1989) *J. Biol. Chem.*, **264**, 14424–14431.
Wang,C.D., Buck,M.A. and Fraser,C.M. (1991) *Mol. Pharmacol.*, **40**, 168–179.
Weiss,M.S., Abele,U., Weckesser,J., Welte,W., Schiltz,E. and Schultz,G.E. (1991) *Science*, **254**, 1627–1630.
Wess,J., Gdula,D. and Brann,M.R. (1991) *EMBO J.*, **10**, 3729–3734.
Wess,J., Nanavati,S., Vogel,Z. and Maggio,R. (1993) *EMBO J.*, **12**, 331–338.
Wong,S.K.-F., Slaughter,C., Ruoho,A.E. and Ross,E.M. (1988) *J. Biol. Chem.*, **263**, 7925–7928.
Zhukovsky,E.A. and Oprian,D.D. (1989) *Science*, **245**, 928–930.

Received on December 2, 1992; revised on January 18, 1993