Protein Science (1999), *8*:969–977. Cambridge University Press. Printed in the USA.

The variable and conserved interfaces of modeled olfactory receptor proteins

YITZHAK PILPEL AND DORON LANCET

Department of Molecular Genetics and Crown Human Genome Center, The Weizmann Institute of Science, Rehovot 76100, Israel

(RECEIVED June 15, 1998; ACCEPTED January 16, 1999)

Abstract

The accumulation of hundreds of olfactory receptor (OR) sequences, along with the recent availability of detailed models of other G-protein-coupled receptors, allows us to analyze the OR amino acid variability patterns in a structural context. A Fourier analysis of 197 multiply aligned olfactory receptor sequences showed an α-helical periodicity in the variability profile. This was particularly pronounced in the more variable transmembranal segments 3, 4, and 5. Rhodopsin-based homology modeling demonstrated that the inferred variable helical faces largely point to the interior of the receptor barrel. We propose that a set of 17 hypervariable residues, which point to the barrel interior and are more extracellularly disposed, constitute the odorant complementarity determining regions. While 12 of these residues coincide with established ligand-binding contact postions in other G-protein-coupled receptors, the rest are suggested to form an olfactory-unique aspect of the binding pocket. Highly conserved olfactory receptor-specific sequence motifs, found in the second and third intracellular loops, may comprise the G-protein recognition epitope. The prediction of olfactory receptor functional sites provides concrete suggestions of site-directed mutagenesis experiments for altering ligand and G-protein specificity.

Keywords: amino acid variability; GPCR; helical moment; olfactory receptors; variable binding site

Most receptor and enzyme families have evolved specificity toward particular endogenous ligand counterparts. As a corollary, the binding sites in such proteins display a high degree of conservation in amino acid sequence. In contrast, protein repertoires that have evolved to cope with a diversity of functional ligands display pronounced binding site sequence variability. The most recognized examples of these are the immune repertoires: immunoglobulins, T-cell receptors, and major histocompatibility complex proteins. In these proteins, the detailed structure of the complementarity determining regions (CDRs) is well established by X-ray crystallography, as well as by functional studies (Branden & Tooze, 1991).

Olfactory receptors (ORs) constitute another large repertoire of proteins geared to recognize diverse ligands. ORs are believed to provide the molecular basis for the recognition of millions of volatile odorous chemicals (Lancet, 1986; Buck & Axel, 1991). OR proteins are G-protein-coupled receptors (GPCRs), coded by a multigene superfamily (Buck & Axel, 1991), estimated to include up to 1,000 genes disposed on clusters on multiple chromosomes (Buck & Axel, 1991; Lancet & Ben-Arie, 1993; Rouquier et al., 1998). It was previously hypothesized, based on protein sequence analyses, that the odorant binding sites are related to transmembrane (TM) segments 3, 4, and 5, which show the highest degree of variability (Buck & Axel, 1991). Yet, detailed information on the identity and spatial disposition of the odorants' contact residues in OR proteins remained unavailable.

Due to the lack of X-ray crystallographic data of GPCRs, their structural modeling (Hibert et al., 1991; Cronet et al., 1993) was initially based on the coordinates of bacteriorhodopsin (Henderson et al., 1990), a non-GPCR seven transmembrane helix protein. These studies, combined with results from site-directed mutagenesis experiments, showed that the ligand-binding sites in the GPCRs reside in the extracellular portion of the helical barrel interior. A bacteriorhodopsin-based model was generated for the rat OR protein OR5, affording a prediction of nine potential ligand-binding residues in TMs 2–7 for the odorant lyral in a docking simulation (Singer & Shepherd, 1994). While the structure of bacteriorhodopsin has the advantage of atomic resolution, it is a less favorable template for GPCRs, as it bears practically no sequence similarity to these proteins, and its two-dimensional map deviates considerably from that of rhodopsin, the light-sensitive GPCR (Scherlter et al., 1993).

More recently, other GPCR models were generated (Baldwin, 1994; Donnelly et al., 1994; Herzyk & Hubbard, 1995; Baldwin et al., 1997), based on a two-dimensional low-resolution map of the rhodopsin molecule (Scherlter et al., 1993). An OR model was generated based on such a rhodopsin two-dimensional map, with

Reprint requests to: Doron Lancet, Department of Molecular Genetics, Genome Center, The Weizmann Institute of Science, Rehovot 76100, Israel; e-mail: bmlancet@weizmann.weizmann.ac.il.

helical orientations optimized to bury in the protein conserved and/or polar amino acids (Afshar et al., 1998). The progress in modeling and functional studies in GPCRs, combined with the recent rapid increase in the number of available OR sequences (Barth et al., 1996; Glusman et al., 1996; Sullivan et al., 1996; Vanderhaeghen et al., 1997), made possible the integration of a comprehensive analysis of the OR sequence variability pattern, along with the prediction of its structural context. We propose here putative OR CDRs, consisting of a set of 17 highly variable residues positioned strategically along one helical face on each of the three variable TM segments 3, 4, and 5.

Results

We generated a global multiple alignment of the deduced amino acid sequences of 197 OR genes, from five species, and a consensus sequence for most of the currently known ORs was identified. We subjected the entire set of ORs to alignment against several other GPCRs (Fig. 1). In all seven TM segments, except for TM6, the alignment between the ORs and other GPCRs was straightforward, due to the presence of highly conserved positions common to the entire GPCR superfamily (marked "G" in Fig. 1). The alignment in these TMs is consistent with the GPCR canonical alignment (Oliveira et al., 1993). Since TM6 in ORs lacks the strong WXP (where x is any amino acid) motif shared by many other GPCRs, we aligned this TM based on a considerably weaker motif, $[F/Y]X_7[F/Y]$, found to be shared by ORs and other GPCRs, such as the adrenergic, dopaminergic, serotonergic, histaminergic, and opioid receptors.

Sequence variability profile and helical periodicity

We computed an amino acid variability profile for the entire alignment of 197 OR sequences (Fig. 2A). Within the TM segments, we observed strong peaks of variability in TM4 and TM5, with smaller peaks also in TM3 and TM6. While there are generally troughs in the variability profile of intracellular loops, in two of the extracellular loops, EC1 and EC3, we saw considerable variability (Fig. 2A). These results quantitatively confirm and extend the previously observed general variability patterns in the OR molecules (Buck & Axel, 1991).

An important question is whether some or all the variable OR residues may line the putative ligand-binding interface. Such interfaces are known to be located in the barrel interior in other GPCRs (Baldwin, 1994; Herzyk & Hubbard, 1995). A structurally relevant clue was obtained through Fourier analysis of the variability profile in the entire OR molecule. From the calculated spectrum (Fig. 3), we can see that the highest peak occurred at a period of 98°, corresponding to 3.67 residues per helical turn, with an α-helical periodicity AP index of 1.94, very near the statistical significance level (Donnelly et al., 1993). This implies an α-helical periodicity in the variability pattern of the entire OR sequence.

To obtain positional information on this periodicity pattern, we calculated the variability AP value in overlapping running windows of 23 amino acids, centered around each position along the entire sequence (Fig. 2B). The results indicated that TM segments 1, 3, 4, 5, and, to a smaller extent, TM segment 7, are characterized by large α-helical variability moments, with significant $AP > 2$ values. Thus, in the three most diversified TM segments 3, 4, and 5, variable positions are inferred to be mainly concentrated along one face of the helix.

The orientation of the variable residues

We then turned to ask what is the relative orientation of the variable helical faces within a three-dimensional framework. For this, we generated homology models for several ORs based on two rhodopsin templates (Herzyk & Hubbard, 1995; Baldwin et al., 1997). After computing the variability moments from the variability profile of each of the seven TM segments, we superimposed this information on the independently predicted structure (Fig. 4). Significantly, we observed that the variability moments of the highly variable TM segments 3, 4, and 5 were largely oriented toward the interior of the receptor barrel, where ligands potentially bind. These predicted orientations were found for both rhodopsin templates (Fig. 4).

The hydrophobicity moments (Eisenberg, 1984) were also computed for each of the helices (Fig. 4). While their AP values were not statistically significant (Donnelly et al., 1993), the cumulative tendency of the hydrophobcity moments was to point toward the lipid bilayer, in additional support for the proposed helical orientations.

The putative odorant CDRs

We turned to localize individual variable residues in the framework of the OR model. We observed that out of 36 hypervariable residues (whose variability value is more than one standard deviation above the average variability in the TM segments), in the entire OR protein, 26 are in the three variable helices, most of which (20 residues) are on the predicted inner surface of the receptor barrel (Fig. 4). Of these 20 amino acid positions, a pronounced majority (17 residues) are located in the extracellular two-thirds of the TM segments, where ligands are know to be bound in other GPCRs. The higher abundance of variable residues on the extracellular portion of the OR TM segments is shown in Figure 2A, whereby several variability peaks display an asymmetry within the TM segments. We propose therefore that the subset of 17 positions constitute the olfactory CDRs (Table 1).

Ligand-contact residues are known for many other GPCRs (Hibert et al., 1991; Baldwin, 1994; Herzyk & Hubbard, 1995). We examined whether our proposed CDR positions are aligned with these functional residues. Based on the alignment of the rhodopsin-like GPCRs to the ORs (Fig. 1), we found that 12 out of the 17 residues of the CDRs, mostly in TM segments 3 and 5, are indeed aligned with identified contact residues in other GPCRs (Fig. 4; Table 1). This overlap lends considerable credibility to our identification of the potential CDR residues of OR proteins. There is, however, a subset of OR variable residues that do not have a clear correspondence to functional positions in other GPCRs. These are mainly concentrated on TM4, lining the cleft between this TM and TM5 (Fig. 4).

Chemical statistics of the variable odorant binding site

The amino acid composition of the potential CDRs, analyzed in all 197 ORs, is depicted in Figure 5A. The distribution is clearly biased toward uncharged amino acids, mostly aliphatic and aromatic, with some weaker preference also seen for polar amino acids, but with a clear paucity of charged amino acids. This distribution may reflect the chemical nature of functional odorants, which usually are mostly hydrophobic compounds with a few polar

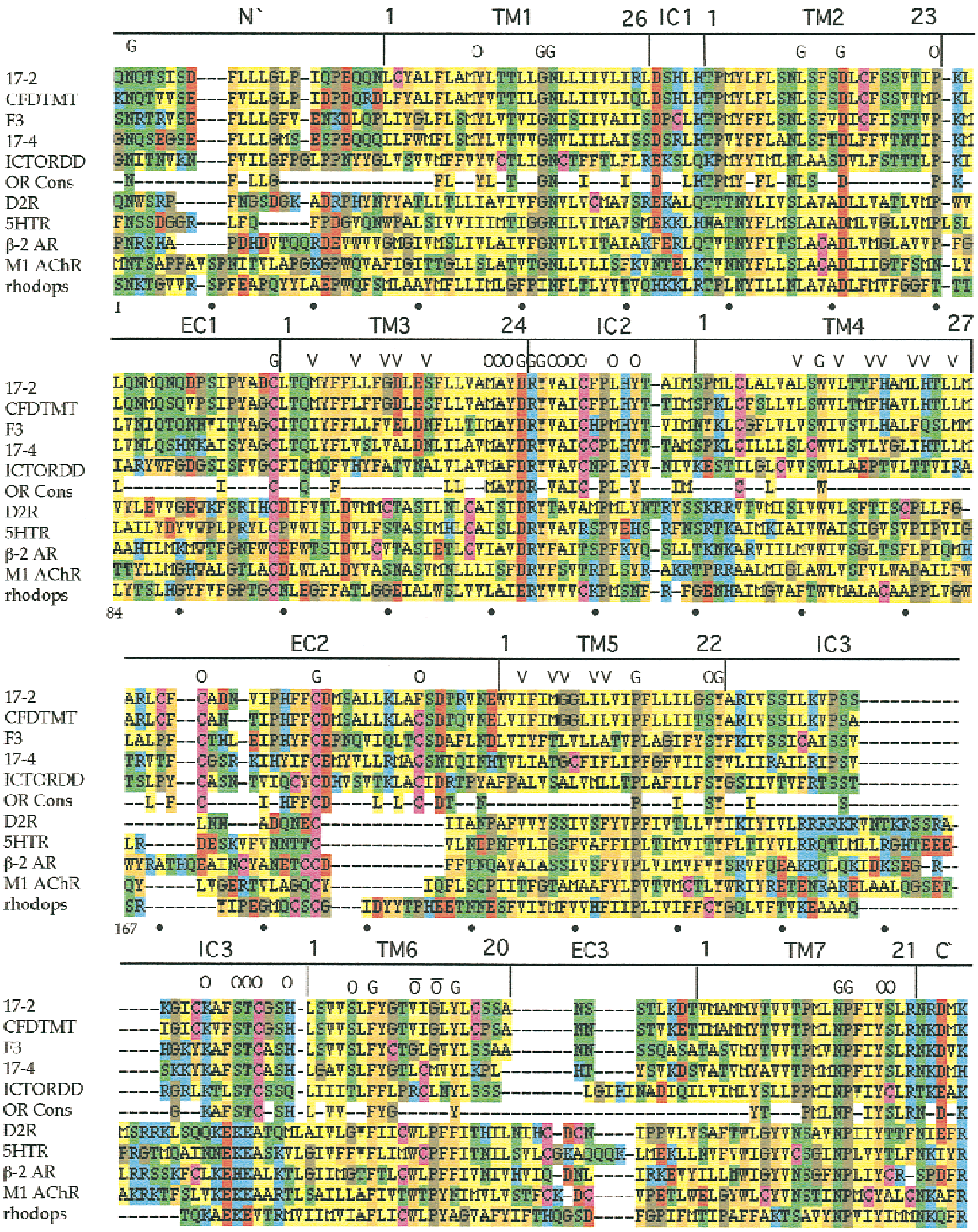

Fig. 1. Multiple alignment of OR proteins (upper rows) and non-OR GPCRs (lower rows). Five typical OR sequences and five non-OR sequences are shown. The row marked "OR Cons" is the consensus of all 197 OR sequences analyzed in this study, calculated by 65% plurality. The OR sequences shown are: 17–2 (human, P30953), ICTORDD (fish, L09217), CFDTMT (dog, P30955), F3 (rat, P23265), and 17–4 (human, P34982). The other GPCR sequences are dopamine D2DR D(2) receptor (human, P14416), serotonin 5H2C receptor (human, P28335), β2 adrenergic receptor (human, P07550), muscarinic M1 receptor (pig, P04761), and rhodopsin (bovine, P02699). The N- and C-termini of all sequences are partially truncated, and the central portion of the long third intracellular loop (between the third and fourth rows) is removed for the dopamine, serotonin, and adrenergic and muscarinic receptors. The boundaries of the seven TM segments and the intracellular and extracellular loops are shown above the sequences. The following positions are marked above the sequences: V, the OR CDR residues (as defined in Fig. 4 and Table 1); G, conserved positions among all GPCRs (Oliveira et al., 1993), which are also highly conserved in ORs; $\bar{O}$, the only two GPCR-conserved positions (in TM segment 6) that do not appear in ORs; O, highly conserved positions unique to ORs (90% plurality). A total alignment position numbering is depicted below the sequence; in addition a TM numbering is given for individual helices.

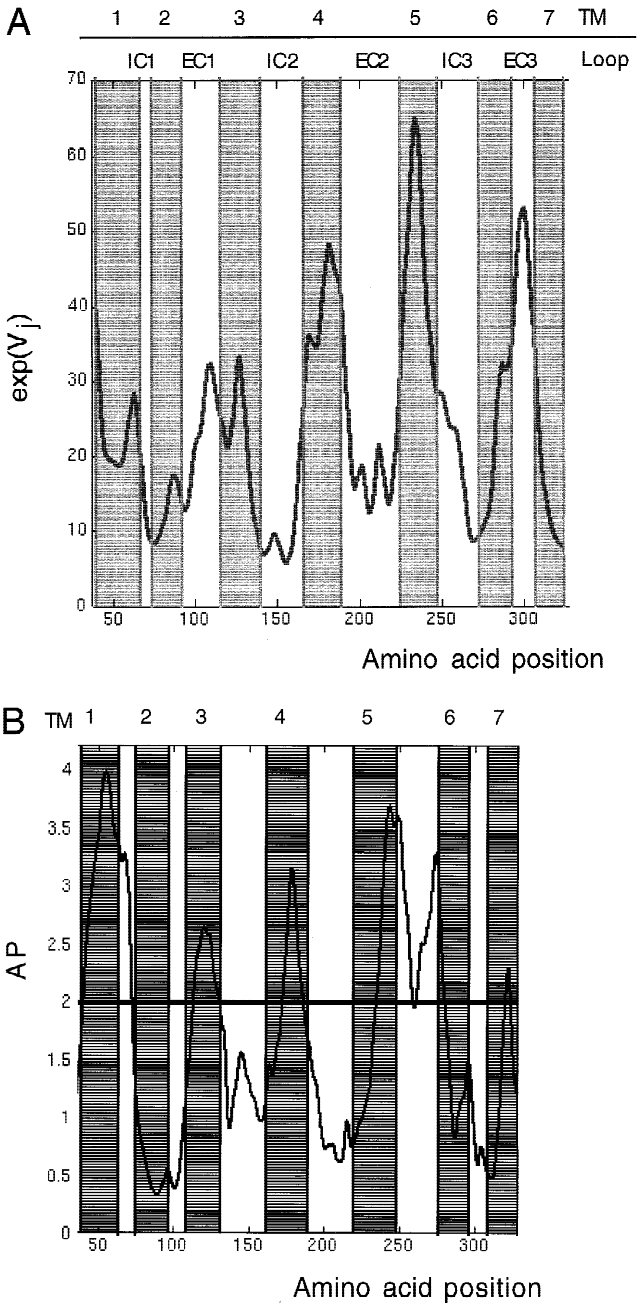

Fig. 2. A: Amino acid variability (V_j) computed along the multiple sequence alignment of all 197 OR sequences using Equation 1 with the BLOSUM65 matrix (Altschul, 1991) as the amino acid similarity matrix *S*. In the profile plotted, $\exp(V_j)$ is shown. Smoothing of the original profile was done using the "hamming" function of the MATLAB/MathWorks Inc. package with a window size = 7. The predicted positions of TM segments, based on the PHDhtm server (Rost et al., 1995), are shown as shaded areas. The N- and C-termini are not shown. **B:** The AP index (Donnelly et al., 1993) calculated in a running window along the alignment. A window of size = 23 amino acid, corresponding to the length of typical TMs, was used. A horizontal line is shown at the significance threshold AP = 2. The profile was smoothed as in **A**.

uncharged functional groups. In contrast, MHC class I molecules, another diverse protein repertoire, revealed a more balanced distribution of the amino acids in the peptide-binding cleft, with no clear bias toward residues with particular properties (Fig. 5B).

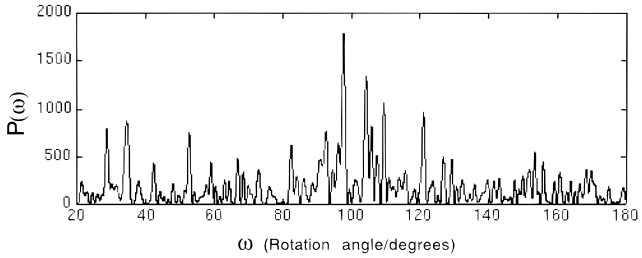

Fig. 3. Fast Fourier Transform analysis of the variability profile. The entire multiple alignment analyzed here yielded an AP value = 1.94, very close to the level of statistically significant value (AP = 2) (Donnelly et al., 1993).

The putative G-protein interface

Signal transduction in OR proteins is assumed to occur by the propagation of structural changes from the odorant CDRs to regions in which sequence motifs are shared among most or all OR

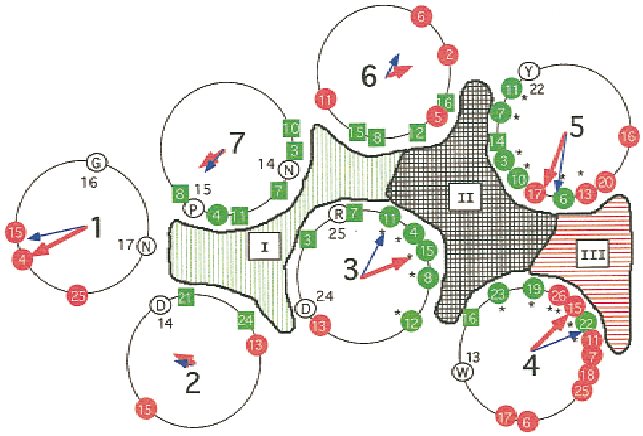

Fig. 4. Schematic two-dimensional representation of the OR seven TM segments, based on the rhodopsin two-dimensional projection map (Scherler et al., 1993). Conserved residues in all the GPCRs (Oliveira et al., 1993) are indicated. In addition, OR positions that align with ligand contact residues in other GPCRs are colored green; OR variable positions that do not align with such residues are colored red. Circles are OR variable positions, while squares are OR conserved positions. The residues in each of the helices are numbered separately, according to the predicted TM boundaries. The 17 residues constituting the putative CDRs are marked with an asterisk. The three marked areas I, II, and III denote the conserved pocket (vertical lines), the variable pocket that corresponds to the ligand-binding pocket in other GPCRs (grid), and the variable pocket that does not correspond to the ligand-binding pocket in other GPCRs (horizontal lines). The OR variability and hydrophobicity (Eisenberg, 1984) α-helical moments are generated from the corrsponding amino acid profiles by assuming a 100° rotation between consecutive amino acids and computing the vectorial sum of moments for all residues. The red thick variability moment arrows and the amino acid positions are depicted according to the rhodopsin template of Herzyk and Hubbard (1995). The blue thin arrows are the variability moments as derived from the coordinates of Baldwin's rhodopsin template. The values for the hydrophobicity moments (not shown), computed as clockwise angles relative to the red arrows are (β) and their AP values are: TM1, β = 110° AP = 0.7; TM2, β = 92° AP = 0.72; TM3, β = 145° AP = 1.1; TM4, β = 214° AP = 0.98; TM5, β = 150° AP = 1.1; TM6, β = 278° AP = 0.68; TM7, β = 88° AP = 0.53; The two-dimensional cross section shown corresponds to the extracellularly disposed section of the barrel—the most relevant region for the CDRs. Due to helical tilts in rhodopsin, the positions and orientations of residues adjacent to the inner leaflet are distorted.

Table 1. *The predicted CDR positions*[a]

OR TM position	Alignment position	Other GPCR amino acid position	Amino acid
TM3 4	103	Vasopressin v1a receptor	K:128
TM3 8	107	Neurokinin-1	V:116
TM3 11	110	Muscarinic m3	Y:148
TM3 12	111	Bovine rhodopsin	E:122
TM3 15	114	Human rhodopsin	L:125
TM4 11	150	NA	
TM4 15	154	NA	
TM4 19	158	NA	
TM4 22	161	Muscarinic m3	P:201
TM4 23	162	Human rhodopsin	P:171
TM4 26	165	NA	
TM5 3	205	Muscarinic m3	T:231
TM5 6	208	Dopamin D1	S:198
TM5 7	209	β-2 adrenergic	S:204
TM5 10	212	β-2 adrenergic	S:207
TM5 11	213	β-2 adrenergic	F:208
TM5 13	215	NA	

[a]The 17 hypervariable CDR positions in the OR proteins (Fig. 4) with their TM and alignment numbering. The other GPCR and GPCR position column show the corresponding functional residues in non-OR GPCRs as derived from the alignment in Figure 1, with their enumeration in the original protein sequence. NA indicates that the variable OR position is not aligned with a GPCR functional residue. Information about functional residues was derived from the GPCR mutant database, GRAP (Kristiansen et al., 1996), and from Baldwin (1994) and Herzyk and Hubbard (1995).

proteins. The TM helices contain certain OR-specific residues that are highly conserved and unique to the OR superfamily (marked "O" in Fig. 1) that could be functionally important. In addition, the variability profile (Fig. 2A) indicates that three of the most highly conserved regions in the OR sequences actually correspond to the intracellular (IC) loops, IC1, IC2, and IC3. All three loops show strong consensus sequences: For IC1 (the "ORIC1" motif), DX_2LHT; for IC2 (the "ORIC2" motif), DR[Y/F]VAICXPLhY; and for the carboxy terminal part of IC3 (the "ORIC3" motif), SaeGRyKAFSTCgSH (with upper case and lower case for >90% and >65% conservation, respectively).

Figure 6A shows a sequence logo (Schneider & Stephens, 1990) of the entire 26 amino acids long IC3, highlighting the ORIC3 motif. The latter contains a conserved cysteine as well as four highly conserved serine/threonine residues interspersed among four positively charged amino acids.

As seen from Figure 2B, the IC3 loop is the only non-TM segment with a Fourier spectrum peak, indicative of an α-helical structure (AP = 2.74). The variability moments have higher magnitudes at the N- and C-termini of this loop, suggesting that it may consist of two helical segments with a short nonhelical connecting stretch. To obtain further support for this proposed structure, we analyzed Fourier spectra for hydropathy and variability separately on the two putative helical segments (Fig. 6). We observed enhanced AP values for the two putative helices, compared to the entire loop. Figure 6B depicts a helical wheel analysis for these helices. It is seen that for the N-terminal helix, the conserved face is hydrophobic, while for the C-terminal helix the conserved positions are mainly polar.

Discussion

The olfactory CDRs

A crucial feature of olfactory receptors is their inherent variability patterns in amino acid sequence, which is necessary for the recognition of the large variety of odorants. In this paper we demonstrate and analyze this variability with respect to the primary, secondary, and tertiary structural levels.

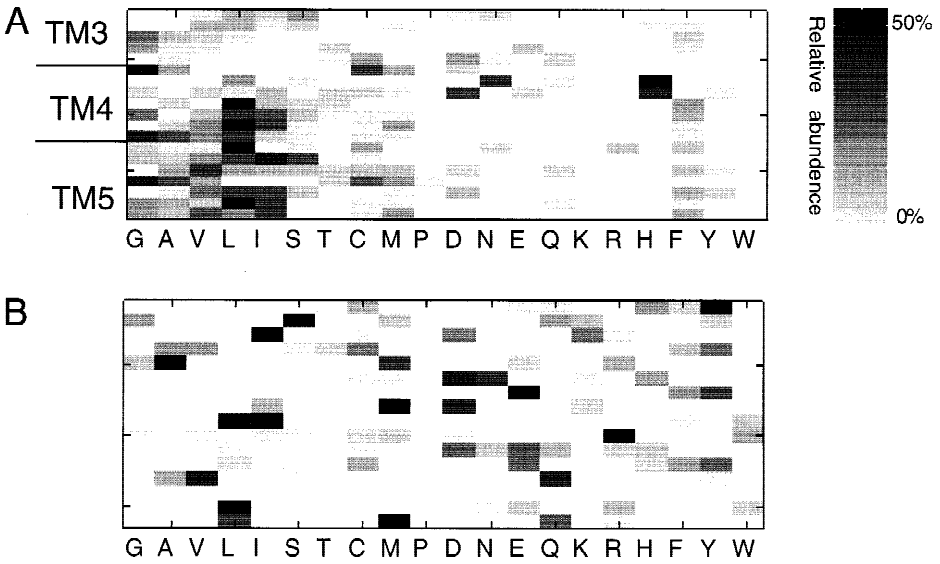

Fig. 5. (**A**) Amino acid composition in the predicted CDRs and (**B**) in the binding site of 150 MHC class I proteins. The frequency of each amino acid is gray-scale coded.

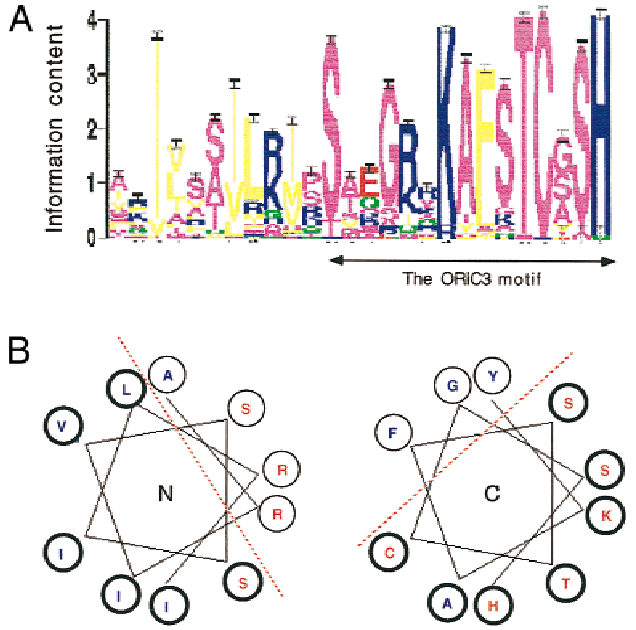

Fig. 6. A: A sequence logo (Schneider & Stephens, 1990) of all 197 ORs at the predicted IC3. The vertical axis represents the information content, i.e., conservation of the amino acids at each position. The ORIC3 motif corresponds to the last 15 positions in this segment. **B:** Helical wheel representation of the first and last 10 amino acids of IC3 loop. The plot was generated using the consensus sequence, with the "helical wheel" program of the GCG package (1994). The AP values in the entire IC3 loop, in the N- and C-terminal portions are 1.85, 3.13, and 2.68, respectively, in the hydrophobicity profile, and 2.74, 2.76, and 2.91, respectively, in the variability profile. The conserved and variable positions are circled with thick and thin lines, respectively, and in blue and red for hydrophobic and hydrophilic positions, respectively.

At the sequence level, we show that a majority of the highly variable residues are present in three of the seven helices, namely TM3, 4, and 5. Previously, about 40 variable residues were identified on these TM segments, based upon analyses of small sets of OR sequences (Buck & Axel, 1991; Rouquier et al., 1998). The present analysis provides a considerably more focused picture of the potential odorant binding pocket. We achieve this by encompassing a much larger number of functional OR sequences and utilizing more quantitative variability measures. We are able, moreover, to highlight a subset of 17 hypervariable residues that fulfill relevant secondary and tertiary structure criteria.

At the secondary structure level, we demonstrate by Fourier analysis that in the three diversified TM segments, the variable residues are clustered on one helical face, as previously seen in a limited analysis of TM5 (Ben-Arie et al., 1993). At the tertiary structure level, based on the homology model, our results suggest that the variable faces of these helices point toward each other, particularly in the extracellular two-thirds of the OR barrel interior. This generates a view of the potential odorant CDRs consistent with the notion that the functional recognition surface should be highly variable. Our results suggest that the ligands should bind at the inner face of the OR transmembrane barrel, as in numerous other GPCRs (Baldwin, 1994; Herzyk & Hubbard, 1995).

Specific odorant-binding residues have been previously predicted by a correlated mutation analysis in TMs 3–5 (Singer et al., 1995a) and by an analysis of nonsynonymous nucleotide substitutions for TM segment 6 (Singer et al., 1996). In addition, functional residues were predicted on the basis of receptor-ligand docking simulations in TMs 2–7 (Singer & Shepherd, 1994; Afshar et al., 1998). These analyses encompassed more limited sets of OR sequences. Of the seven contact residues predicted by these studies, only 2 are in overlap to the set of 17 hypervariable residues identified here. The rest of the previously predicted contact residues appear to represent more conserved positions in the context of the larger number of OR sequences analyzed here. This, however, does not preclude their potential involvement in ligand contacts.

In general, it is possible that some ligand interactions will be mediated by conserved amino acids (Singer & Shepherd, 1994). As an example, while the TM6 helix shows low variability in ORs, it is clearly important in ligand binding by other GPCRs, and could perhaps take part in odorant binding as well. TM6 has a set of three inward-facing aromatic residues, which includes the $[F/Y]X_7[F/Y]$ motif in both ORs and other GPCRs. These residues have been proposed to interact with aromatic rings in the neurotransmitter ligands (Hibert et al., 1991) and could similarly interact with aromatic moieties of odorant molecules.

In the present homology model, the inward orientation seen for most variable residues is a direct consequence of the alignment to rhodopsin and other GPCRs. A different strategy was also used to construct an OR model (Afshar et al., 1998), in which the helix orientations are based on the assumption that specific, highly conserved, or polar residues face the barrel interior, as suggested for GPCR families who share ligand specificity (Taylor et al., 1994; Alkorta & Loew, 1996). Notably, the two OR models appear to agree well with respect to the orientation of the more conserved TM segments 1, 2, 6, and 7, the three variable TM segments, particularly in the orientation of TM3.

Comparisons to other GPCRs

The homology model and the multiple alignment suggest that the olfactory receptor CDRs partially coincide with the regions in which ligand contact residues have been localized in many other GPCRs. Specifically, most of the hypervariable amino acids in TM segments 3 and 5 (lining the barrel region marked II in Fig. 4) are exactly aligned with GPCR ligand-contact residues. This lends strong support to their present proposed functional assignment. In contrast, in the region of the cleft encompassed by TM segments 2, 6, and 7 (marked I in Fig. 4), where a large number of GPCR ligand-contact residues are found, practically no OR variable residues occur.

Other hypervariable residues of the modeled olfactory CDRs, which do not correspond to GPCR binding sites, are located in the cleft between TMs 4 and 5 (marked III in Fig. 4) and are partially facing the lipid phase. It is possible that some of the sequence variability is the result of a random genetic drift and is unrelated to odorant binding. Yet, the observed dense clustering of hypervariable residues lends support to the notion that many of them play a role in odorant recognition. If true, this more exposed interface would represent an OR-unique attribute. Such a CDR, partially facing the lipid phase, could perhaps be important for binding the more hydrophobic odorants, which may reach this site via the lipid bilayer. This mechanism may be akin to the "hydrophobic vacuum cleaner" model for the multi drug resistance (MDR) pumps (Bolhuis et al., 1997). Interestingly, many of the variable residues in

this more peripheral aspect of the CDR are apolar, while a considerable fraction of the variability in the central core relates to polar and rarely even charged residues.

In the present model, based on two rhodopsin templates (Herzyk & Hubbard, 1995; Baldwin et al., 1997), TM4 is oriented such that Trp161 (marked 13 in Fig. 4) is exposed to the lipid phase, as also corroborated by chemical modification studies (Davison & Findlay, 1986), potentially facing TM2. This, however, is inconsistent with previous GPCR models (Donnelly et al., 1989) in which this amino acid was suggested to face the TM3 segment, more toward the protein interior, a suggestion based on its high conservation in the entire superfamily. Such orientation, if applied to the ORs, would imply that the variability moment in TM4 should face the lipid phase with practically no variable positions in the barrel interior. Additional work will be needed to determine the exact orientation of this TM segment.

Considerable variability is also evident in two of the extracellular loops, EC1 and EC3. The possibility cannot be excluded that some odorants actually interact with these loops, as in the case of other GPCRs (Baldwin, 1994). Alternatively, other functions related to receptor specificity might be mediated by extracellular variable loops indicating a potential role in specific axonal guidance (Singer et al., 1995b).

The G-protein binding site

The G-protein interface of GPCRs has been thoroughly studied with both the second and third cytoplasmic loops implicated in this function (Hedin et al., 1993). Specific residues in these loops have been shown to be essential for the interaction with specific G-protein types, in particular positively charged residues. However, due to the diversity of G-protein heterotrimers involved in the interaction with the variety of all known GPCRs, a very clear relevant amino acid consensus has not emerged. The present analysis of OR protein sequences provides a unique opportunity in this respect: the availability of a very large group of receptors, all presumably sharing an interaction with the same G-protein (Pace & Lancet, 1986; Jones & Reed, 1989). Our analysis shows a strong consensus sequence at the first and second intracellular loops and at the C-terminal end of the third intracellular loop (the ORIC1, ORIC2, and ORIC3 motifs). Both the ORIC2 and the ORIC3 consensus sequences are found to be unique to ORs, as subjecting them to BLAST search (Altschul et al., 1990) against the nonredundent protein sequence database yielded only OR sequences. Further, all ORs analyzed in this paper are identified in such a search with a minimum of 65% sequence identity. It is thus likely that our analysis has resulted in at least a partial definition of the G-protein interface of the ORs. The ORIC3 consensus contains a strongly conserved motif, KXXST, potentially involved in kinase-mediated serine/threonine phosphorylation. Such process is involved in the desensitization of some GPCRs (Probst et al., 1992), including OR proteins (Boekhoff et al., 1997).

Amino acid hydropathy analysis in the β-adrenergic receptor (Strader et al., 1989) and random mutagenesis analyses in the muscarinic m5 receptor (Eubanks et al., 1996) predicted that both ends of the IC3 loop might adopt an α-helical secondary structure. Based on the variability and hydropathy moment analyses done here, we suggest, similarly, that in OR proteins the considerably shorter IC3 loop is composed of two α-helices of 10 amino acids each, connected by a short stretch of 6 amino acids. The amino and carboxy terminal IC3 helical structures may, in fact, be considered as respective intracellular extensions of TM segments 5 and 6, as is the case for bacteriorhodopsin (Pebay-Peyroula et al., 1997) and rhodopsin (Baldwin et al., 1997).

Limitations of the modeling procedure

Most of the foregoing insight is made possible by applying a homology modeling approach, using bovine rhodopsin models as templates. This procedure suffers from several aspects of inherent inaccuracy. The overall sequence identity between the currently known ORs and bovine rhodopsin is 21%, below the threshold required for safe homology modeling, at least 30% (Sander & Schneider, 1991), as established for globular proteins. However, the well-defined conserved amino acid positions found in all TM segments, except TM6, in addition to the assumption that gaps are very rare in TM segments, allow the generation of a reliable alignment, on which the homology modeling is based. The lack of confidence in the alignment of TM6 has only a minor effect on our functional predictions because this TM segment does not appear to belong to the CDRs. An additional point of weakness is that all current models for rhodopsin are based on low resolution experimental data and are thus not expected to be accurate at the atomic level. However, the functional conclusions drawn here are mainly related to TM helix orientations, not to specific coordinates of individual amino acids, thus requiring only a relatively low spatial resolution and accuracy.

Despite limitations, the present OR model provides a general prediction of functional odorant contact residues. Successful binding experiments have recently led to assignment of odorant specificities to particular ORs (Raming et al., 1993; Wellerdieck et al., 1997; Zhao et al., 1998). The model proposed here provides rational guidelines for site-directed mutagenesis experiments, which in turn will provide experimental tests and refinements of the model.

Materials and methods

OR sequences

We analyzed a collection of 197 complete and partial OR sequences from five species as detailed in URL: http://bioinfo.weizmann.ac.il/~bnpilpel/ORseq. We obtained the amino acid sequences either from the SWISSPROT database or we conceptually translated them from GenBank.

Multiple sequence alignments

We performed multiple alignment of the deduced amino acid sequences of all ORs by a fully automated application of the Clustal W program (Thompson et al., 1994). We applied the default pairwise gap opening penalty of 10 and extension penalty of 0.05. We did segmental alignments of ORs with other GPCRs with the Clustal W program using a procedure that allows segment-specific gap penalties. For the TM segments, we used a higher gap opening penalty of 20 in order to prevent insertions. The interhelical loops were aligned with the default penalty parameters. In rare cases the alignments were manually edited in TM segments to remove gaps and to ensure that conserved positions were aligned.

Amino acid variability profiles

We calculated the amino acid variability value V_j at alignment position j as

$$V_j = \sum_{i=1}^{20} \sum_{k=i}^{20} P_{ij} P_{kj} / S_{ik} \quad (1)$$

where P_{ij} is the fraction of sequences with amino acid i at position j of the alignment, and S_{ik} is the ik element in an amino acid similarity matrix such as PAM (Dayhoff, 1978) or BLOSUM (Altschul, 1991). This variability measure and the classical entropy function, previously used for measuring sequence variability (Sander & Schneider, 1991) are mathematically related and were recently shown to be two special cases of a more general term (Baczkowski et al., 1997).

Fourier analysis and helical periodicity detection

To assess the dominant periods that characterize profiles of the hydrophobicity and the variability of the analyzed sequences, we used the Fast Fourier Transform (FFT) algorithm, implemented in the "FFT" function of the Matlab/MathWorks package. Periodicity detection enhancement and tests for statistical significance, using the alpha helical periodicity AP index, were done as described (Donnelly et al., 1993).

Prediction of TM segments

We predicted the number, boundaries, and topology of the TM segments using the neural network algorithm implemented in the "PHD" server (Rost et al., 1995), with the PHDhtm and PHD topology programs.

Homology modeling

We constructed three-dimensional homology models for OR proteins, based on bovine rhodopsin as a template, using the automated SwissModel server (Peitsch et al., 1996) for the all-atoms model (Herzyk & Hubbard, 1995). We used the "homology" and "discover" modeling modules of the Biosym/MSI package, with minimization parameters as described (Cronet et al., 1993), modeling the ORs based on the α-carbon rhodopsin model (Baldwin et al., 1997). With this template, we generated automatically the side-chain conformations with the biopolymer module of the MSI package. In both cases we used the proposed alignment in the TM segments between rhodopsin and the ORs (Fig. 1).

All the alignment and modeling data are publicly available on http://bioinfo.weizmann.ac.il/~bnpilpel/msa and http://bioinfo.weizmann.ac.il/~bnpilpel/ORModel.

Acknowledgments

This research was supported by grants from US National Institute of Health (DC00305), and a Wolfson Research Award of the Israel Academy of Science, the BMFT and Infrastructure grants of the Israel Ministry of Sciences and the Arts, and the Gesellschaft für Biotechnologische Forschung, Braunschweig. We thank Ephraim Katchalsky-Katzir, Michael Levitt, David de Graaf, Dror Sharon, and Shai Rosenwald for helpful discussions. We thank Daniel Segré for help with Fourier analysis, Gustavo Glusman for help with sequence analysis.

References

Afshar M, Hubbard RE, Demaille J. 1998. Towards structural models of molecular recognition in olfactory receptors. *Biochimie 80*:129–135.
Alkorta I, Loew GH. 1996. A three-dimensional model of the delta opioid receptor and ligand-receptor complexes. *Protein Eng 9*:573–583.
Altschul S. 1991. Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol 219*:555–565.
Altschul SF, Gish W, Miller W, Myers E, Lipman D. 1990. Basic local alignment search tool. *J Mol Biol 215*:403–410.
Baczkowski A, Joanes D, Shamia G. 1997. Properties of a generalized diversity index. *J Theor Biol 188*:207–213.
Baldwin J. 1994. Structure and function of receptors coupled to G-proteins. *Curr Opin Cell Biol 6*:180–190.
Baldwin J, Schertler G, Unger V. 1997. An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. *J Mol Biol 272*:144–164.
Barth A, Justice N, Ngai J. 1996. Asynchronous onset of odorant receptor expression in the developing zebrafish olfactory system. *Neuron 16*:23–34.
Ben-Arie N, Lancet D, Taylor C, Khen M, Walker N, Ledbetter D, Carrozzo R, Pate K, Sheer D, Lehrach H. 1993. Olfactory receptor gene cluster on human chromosom 17: Possible duplication of an ancestral receptor repertoire. *Hum Mol Genet 3*:229–235.
Boekhoff I, Touhara K, Danner S, Inglese J, Lohse M, Breer H, Lefkowitz R. 1997. Phosducin, potential role in modulation of olfactory signaling. *J Biol Chem 272*:4606–4612.
Bolhuis H, van Veen H, Poolman B, Driessen A, Konings W. 1997. Mechanisms of multidrug transporters. *FEMS Microbiol Rev 21*:55–84.
Branden C, Tooze J. 1991. *Introduction to protein structure*. New York: Garland Publishing Inc.
Buck L, Axel R. 1991. A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. *Cell 65*:175–187.
Cronet P, Sander C, Vriend G. 1993. Modeling the transmembrane seven helix bundle. *Protein Eng 6*:59–64.
Davison M, Findlay J. 1986. Identification of the sites in opsin modified by photo-activated azido[^{125}]iodobenzene. *Biochem J 236*:389–395.
Dayhoff M. 1978. *Atlas of protein sequence and structure*. Washington: National Biomedical Research Foundation, pp 345–352.
Donnelly D, Findlay J, Blundell T. 1994. The evolution and structure of aminergic G-protein-coupled receptors. *Receptors Channels 2*:61–78.
Donnelly D, Johnson M, Blundell T, Saunders J. 1989. An analysis of the periodicity of conserved residues in sequence alignment of G-protein-coupled receptors. *FEBS Lett 251*:109–116.
Donnelly D, Overington J, Ruffle S, Nugent J, Blundel T. 1993. Modeling α-helical transmembrane domains: The calculation and use of substitution tables for lipid facing residues. *Protein Sci 2*:55–70.
Eisenberg D. 1984. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc Nat Acad Sci 81*:140–144.
Eubanks D, Burstein E, Spalding T, Bräuner-Osborne H, Brann M. 1996. Structure of a G-protein-coupling domain of a muscarinic receptor predicted by random saturation mutagenesis. *J Biol Chem 271*:3058–3065.
GCG. 1994. Program manual for the Wisconsin package, version 8. Madison, Wisconsin: Genetics Computer Group.
Glusman G, Clifton S, Roe R, Lancet D. 1996. Sequence analysis in the olfactory receptor gene cluster on human chromosome 17: Recombinatorial events affecting receptor diversity. *Genomics 37*:147–160.
Hedin KE, Duerson K, Claphan DE. 1993. Specificity of receptor-G protein interactions: Searching for the structure behind the signal. *Cell Signal 5*:505–518.
Henderson R, Baldwin JM, Ceska TA, Zemlin F, Beckmann E, Downing KH. 1990. Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J Mol Biol 213*:899–929.
Herzyk P, Hubbard RE. 1995. Automated method for modeling seven-helix transmembranal receptors from experimental data. *Biophysical J 69*:2419–2442.
Hibert MF, Trumpp-Kallmeyer S, Bruinvels A, Hoflack J. 1991. Three-dimensional models of neurotransmitter G-binding protein-coupled receptors. *Mol Pharmacol 40*:8–15.
Jones DT, Reed RR. 1989. Golf: An olfactory neuron specific-G-protein involved in odorant signal transduction. *Science 244*:790–795.
Kristiansen K, Dahl SG, Edvardsen O. 1996. A database of mutants and effects of site-directed mutagenesis experiments on G-protein-coupled receptors. *Proteins 26*:81–94.
Lancet D. 1986. Vertebrate olfactory reception. *Ann Rev Neurosci 9*:329–355.
Lancet D, Ben-Arie N. 1993. Olfactory receptors. *Current Biol 3*:668–674.
Oliveira L, Paiva ACM, Vriend G. 1993. A common motif in G-protein-coupled seven transmembrane helix receptors. *J Comput Aided Mol Design 7*:649–658.

Pace U, Lancet D. 1986. Olfactory GTP-binding protein: Signal-transducing polypeptide of vertebrate chemosensory neurons. *Proc Natl Acad Sci USA 83*:4947–4951.

Pebay-Peyroula E, Rummel G, Rosenbusch JP, Landau E. 1997. X-ray structure of bacteriorhodopsin at 2.5 Å from microcrystals grown in lipid cubic phases. *Nature 277*:1676–1681.

Peitsch MC, Herzyk P, Wells TN, Hubbard RE. 1996. Automated modeling of the transmembrane region of G-protein-coupled receptor by Swiss-model. *Receptors Channels 4*:161–164.

Probst WC, Snyder LA, Schuster DI, Brosius J, Sealfon SC. 1992. Sequence alignment of the G-protein-coupled receptor superfamily. *DNA Cell Biol 11*:1–20.

Raming K, Krieger J, Strotmann J, Boekhoff I, Kubick S, Baumstark C, Breer H. 1993. Cloning and expression of odorant receptors. *Nature 361*:353–356.

Rost B, Casadio R, Fariselli P, Sander C. 1995. Transmembrane helices predicted at 95% accuracy. *Protein Sci 4*:521–533.

Rouquier S, Taviaux S, Trask BJ, Brand-Arpon V, van den Engh G, Demaille J, Giorgi D. 1998. Distribution of olfactory receptor genes in the human genome. *Nat Genet 18*:243–250.

Sander C, Schneider R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins 9*:56–68.

Scherlter GFX, Villa C, Henderson R. 1993. Projection structure of rhodopsin. *Nature 362*:770–772.

Schneider TD, Stephens RM. 1990. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res 18*:6097–6100.

Singer M, Oliveira L, Vriend G, Shepherd GM. 1995a. Potential ligand-binding residue in rat olfactory receptors identified by correlated mutation analysis. *Receptors Channels 3*:89–95.

Singer MS, Shepherd GM. 1994. Molecular modeling of ligand-receptor interactions in the OR5 olfactory receptor. *Neuroreport 5*:1297–3000.

Singer MS, Shepherd GM, Greer CA. 1995b. Olfactory receptors guide axons. *Nature 377*:19–20.

Singer MS, Weisinger-Lewin Y, Lancet D, Shepherd GM. 1996. Positive selection moments identify potential functional residues in human olfactory receptors. *Receptors Channels 4*:141–147.

Strader CD, Sigal IS, Dixon RAF. 1989. Structural basis of β-adrenergic receptor function. *FASEB 3*:1825–1832.

Sullivan SL, Adamson MC, Ressler KJ, Kozak CA, Buck L. 1996. The chromosomal distribution of odorant receptor genes. *Proc Natl Acad Sci 93*: 884–888.

Taylor WR, Jones DT, Green NM. 1994. A method for α-helical integral membrane protein fold prediction. *Proteins 18*:281–294.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties, and weight matrix choice. *Nucleic Acids Res 22*:4673–4680.

Vanderhaeghen P, Schurmans S, Vassart G, Parmentier M. 1997. Specific repertoire of olfactory receptor genes in the male germ cells of several mammalian species. *Genomics 39*:239–246.

Wellerdieck C, Oles M, Pott L, Korsching S, Gisselmann G, Hatt H. 1997. Functional expression of odorant receptors of the zebrafish *Danio rerio* and of the nematode *C. elegans* in HEK293 cells. *Chem Senses 22*:467–476.

Zhao H, Ivic L, Otaki JM, Hashimoto M, Mikoshiba K, Firestein S. 1998. Functional expression of a mammalian odorant receptor. *Science 279*:237–242.