Protein Science (1993), *2*, 55–70. Cambridge University Press. Printed in the USA.

Modeling α-helical transmembrane domains: The calculation and use of substitution tables for lipid-facing residues

DAN DONNELLY,[1,3] JOHN P. OVERINGTON,[1,4] STUART V. RUFFLE,[2] JONATHAN H.A. NUGENT,[2] AND TOM L. BLUNDELL[1]

[1] ICRF Unit of Structural Molecular Biology, Department of Crystallography, Birkbeck College, Malet St., London WC1E 7HX, United Kingdom
[2] Department of Biology, Darwin Building, University College London, Gower St., London WC1E 6BT, United Kingdom

(RECEIVED June 9, 1992; REVISED MANUSCRIPT RECEIVED September 28, 1992)

Abstract

Amino acid substitution tables are calculated for residues in membrane proteins where the side chain is accessible to the lipid. The analysis is based upon the knowledge of the three-dimensional structures of two homologous bacterial photosynthetic reaction centers and alignments of their sequences with the sequences of related proteins. The patterns of residue substitutions show that the lipid-accessible residues are less conserved and have distinctly different substitution patterns from the inaccessible residues in water-soluble proteins. The observed substitutions obtained from sequence alignments of transmembrane regions (identified from, e.g., hydrophobicity analysis) can be compared with the patterns derived from the substitution tables to predict the accessibility of residues to the lipid. A Fourier transform method, similar to that used for the calculation of a hydrophobic moment, is used to detect periodicity in the predicted accessibility that is compatible with the presence of an α-helix. If the putative transmembrane region is identified as helical, then the buried and exposed faces can be discriminated. The presence of charged residues on the lipid-exposed face can help to identify the regions that are in contact with the polar environment on the borders of the bilayer, and the construction of a meaningful three-dimensional model is then possible. This method is tested on an alignment of bacteriorhodopsin and two related sequences for which there are structural data at near atomic resolution.

Keywords: Fourier transform; lipid-accessible side chains; periodicity; secondary structure prediction; substitution tables; transmembrane helices

We remain unable to predict the three-dimensional structure of a protein from its amino acid sequence, even with the help of the knowledge gained from an increasing number of experimentally determined structures. The prediction of the structure of integral membrane proteins may seem to be even further from our grasp because there are still very few known structures for this class of proteins. However, membranes are essentially two-dimensional, and as such they provide a powerful constraint upon the arrangement of the elements that cross them. These elements are often α-helices where the need to form hydrogen bonds from all main-chain -NH and -CO functions is easily satisfied (Engelman et al., 1986). There are fewer ways that this can be achieved for β-strands, though of course membrane-buried β-structures do exist (e.g., Wallace, 1990; Weiss et al., 1991). Furthermore, transmembrane helices are likely to be approximately perpendicular to the plane of the membrane and will therefore pack together in a parallel or antiparallel fashion (this again will not always be the case as is evident from the work of Kühlbrandt & Wang [1991]).

Reprint requests to: Tom L. Blundell, ICRF Unit of Structural Molecular Biology, Department of Crystallography, Birkbeck College, Malet St., London WC1E 7HX, United Kingdom.
[3] Present address: Department of Biochemistry and Molecular Biology, The University of Leeds, Leeds LS2 9JT, United Kingdom.
[4] Present address: Discovery Chemistry, Pfizer Limited, Ramsgate Rd., Sandwich, Kent CT13 9NS, United Kingdom.

The structure prediction of α-helical membrane proteins can often be viewed as a two-dimensional problem for which four pieces of information are required:

1. The regions of sequence that form the transmembrane helices;
2. The basic topology of the transmembrane domain;
3. The side of each helix that faces the interior of the helical bundle;
4. The relative depth that each helix is inserted into the membrane.

The transmembrane regions can be identified from the amino acid sequence by hydrophobicity and hydropathy analysis or by proteolytic cleavage and chemical probe methods (e.g., Jennings, 1989, for a review). The prediction of the topology is more difficult, but because the helices are likely to pack in a parallel or an antiparallel manner, their possible arrangement is limited once the number and the relative directions of the helices are known. The number of candidates can be reduced further using approaches such as that of Engelman et al. (1980). In this paper we address the third and fourth requirements. This information makes it possible to construct three-dimensional models and to predict the residues on different helices that may mutually interact.

The structures of the photosynthetic reaction centers provide the first high-resolution examples of integral membrane protein structures. They include proteins from both *Rhodopseudomonas viridis* (1PRC, Brookhaven Protein Data Bank; Bernstein et al., 1977) and *Rhodobacter sphaeroides* (1RCR) (Deisenhofer et al., 1984, 1985; Chang et al., 1986; Allen et al., 1987a,b, 1988; Yeates et al., 1987, 1988; Komiya et al., 1988), which are both composed of three protein subunits: L, M, and H. The L and M subunits both contain five transmembrane regions and share a similar tertiary fold. They show sequence homology with other photosynthetic bacterial reaction centers, as well as with the D1 and D2 subunits of the photosystem II complex of cyanobacteria, algae, and green plants (Deisenhofer et al., 1985). The H subunit has one membrane-spanning helix, but there is no equivalent subunit in photosystem II.

Comparisons between the reaction center structures and those of water-soluble proteins (Rees et al., 1989b) have shown the expected difference in surface polarity brought about by the large difference in their surrounding environments. Despite this, the atomic packing and surface area are similar in both classes of protein. Another feature common to both membrane and water-soluble proteins is that the surface residues are less conserved than those in the interior (Smith, 1968; Chothia & Lesk, 1986; Komiya et al., 1988). The comparison therefore suggests that, although the surfaces of aqueous and membrane proteins differ in polarity, their interior structure is similar (Rees et al., 1989b). However, this is probably not the case for all membrane proteins, especially some ion channels that require very polar centers.

The differences between the substitution patterns of surface and buried residues have been described for water-soluble proteins (Overington et al., 1990, 1992). In this paper we use alignments of sequences from proteins that are homologous to the bacterial photoreaction center structures, in order to calculate substitution tables for those residues that are accessible to the lipid fraction of the bilayer. We make the assumption that the mutational properties of the internal residues in membrane proteins will be similar to those calculated for aqueous proteins (Overington et al., 1990, 1992). This is to avoid basing the substitution tables for buried residues only upon the reaction center fold, where many inaccessible residues have specific structural and functional roles—for example binding cofactors—unique to the fold and function of this particular protein family. This assumption seems reasonable based upon the comparison described in Rees et al. (1989b). We compare the substitution patterns of lipid-accessible residues with those of buried and aqueous-accessible residues in water-soluble proteins.

Previous workers have used Fourier transform methods to identify the periodic differences of hydrophobic and hydrophilic residues in order to identify amphipathic helices from sequences and sequence alignments (Eisenberg et al., 1984; Cornette et al., 1987; Bowie et al., 1990). However, the difference in the hydrophobicity of buried and exposed residues is less in membrane proteins compared to water-soluble proteins (Rees et al., 1989a), and therefore this method is less successful. Because buried residues are more conserved than exposed residues in both protein classes (Smith, 1968; Chothia & Lesk, 1986; Komiya et al., 1988) the periodicity of conserved/variable residues can also be used to predict the presence of helices (Komiya et al., 1988; Donnelly et al., 1989; Rees et al., 1989a). In this method (Komiya et al., 1988), a variability profile (V) is calculated from a sequence alignment and used to calculate ψ, an index that is a measure of the helical periodicity in the profile V. The V_j elements of this profile are defined by the number of different residue types at each position j in the alignment. This provides a method that is independent of the hydrophobicity procedure.

We use the substitution tables for buried residues (Overington et al., 1990, 1992) and for lipid-accessible residues (described below) to predict the orientation of the seven helices in bacteriorhodopsin from an alignment of three sequences. This is achieved using a modified version of the standard Fourier transform method (described below). The buried face of each helix is identified, and this information is then used to predict the point at which the helix makes contact with the aqueous environment at the borders of the bilayer. The results are compared with the structure of bacteriorhodopsin (1BRD; Henderson et al., 1990) and also with the method described in Komiya

et al. (1988) that we have modified to take into account the effects of conserved lipid-facing proline residues.

Methods

Construction of substitution tables

The two L and two M subunits from the reaction centers from *R. viridis* and *R. sphaeroides* were aligned using the alignment program COMPARER (Šali & Blundell, 1990), which takes into account features of the three-dimensional structures of the proteins being compared. The aligned proteins showed between 28 and 58% pairwise sequence identity. We aligned the regions of the transmembrane helices that are within the lipid portion of the bilayer (Yeates et al., 1987; Rees et al., 1989b) with the equivalent sequences of five bacterial and cyanobacterial reaction center sequences and also with 28 sequences from the D1 and D2 subunits from photosystem II in green plants (sequences extracted from the OWL database; Bleasby & Wootton, 1990). Identical sequences within each of the five alignments were removed. The single lipid-spanning regions of the H subunits from the two structures were aligned to give a sixth alignment. The alignment for helix 2 is shown in Figure 1.

```
wnrfms          WliAsffMfvAVwsWwgR
rcem$rhovi      WlmAGlfMtlSLgswwiR
wnrfls          WqiITiCAtgAFvsWalR
rcel$rhovi      WqaItvCALgAFiSWmlR

rcem$rhoca      WQIASLFMAISVIAWWVR
rcem$rhocu      WQIAGFFLTTSILLWWVR
rcem$chlau      WLIATFFLTVSIFAWYMH
rcel$rhocu      WQIITFSAIGAFVSWALR
rcel$chlau      WQMTVLFATIAFFGWMMR
psba$horvu      YELIVLHFLLGVACYMGR
psba$chlmo      YQLIVCHFFIGICCYMGR
psba$chlre      YQLIVCHFLLGVYCYMGR
psba$cyapa      YQFVVMHFLLGVACYMGR
psb2$synp7      YQLVVFHFLIGVFCYMGR
psb1$synp7      YQLVVFHFLLGISCYMGR
psb2$syny3      YQLVVFHFLIGIFCYMGR
psba$fredi      YQLVIFHFLLGCACYLGR
psba$euggr      YQLIVCHFFIGICSYMGR
psb1$syny3      YQLNVFHFLIGIFCYLGR
f2pmd2          WTFVALHGAFGLIGFMLR
f2spd2          WAFVALHGAFALIGFMLR
f2rzd2          WTFVALHGAFALIGFMLR
psbd$chlre      WAFVALHGAFGLIGFMLR
psbd$synp7      WNFVALHGAFALIGFMLR
```

Fig. 1. Alignment of the sequences of helix 2 from the four structures with the equivalent regions of 20 related sequences. The lipid-accessible residues in the four structures are shown in lowercase. The codes for identifying the sequences are those from the OWL protein sequence database (Bleasby & Wootton, 1990).

The percentage side-chain accessibility (*a*: Lee & Richards, 1971) was calculated for all residues in both structures from the intact L, M, and H complex, including the cofactors. Those residues within the lipid region of the bilayer with *a* greater than 7% (Hubbard & Blundell, 1987) were considered as lipid accessible. The sequence of each one of the four subunit structures was compared, in a pairwise fashion, with each other structure and sequence in the alignment. The substitutions observed for the lipid-accessible residues were scored in a 20 × 20 matrix F^l composed of elements f^l_{ik} that represent the frequency of substitutions of the lipid-accessible residue type k to residue type i; 3,853 residue substitutions were observed.

This frequency matrix was converted to a probability matrix by dividing the frequency of each substitution by the total number of substitutions observed for that particular residue type.

$$p^l_{ik} = f^l_{ik} \Big/ \sum_{i=1}^{20} f^l_{ik}. \quad (1)$$

The probability matrix P^l is composed of elements p^l_{ik} representing the probability of the substitution of residue type k to i. Standard errors were calculated as $\sqrt{x(n-x)/n^3}$ (where x = the frequency of occurrence of an event and n = sample size) to give errors that correspond to one standard deviation.

The difference between the substitution patterns of lipid-accessible residues and those residues buried in the cores of water-soluble proteins (P^b, from Overington et al. [1992]) can be readily observed by calculating the difference matrix (D^{bl}) (Table 1).

$$d^{bl}_{ik} = p^b_{ik} - p^l_{ik}. \quad (2)$$

A difference probability greater than 0 indicates that the substitution (from a residue k to residue i) is more likely if residue type k is buried, whereas a value less than 0 indicates a substitution more likely at a position exposed to lipid.

The standard Fourier transform procedure

A property profile U is calculated so that a property U_j is assigned at each position j in a sequence or sequence alignment over a window size N. The moment M can be calculated as

$$M = \left\{ \left[\sum_{j=1}^{N} U_j \sin(j\omega) \right]^2 + \left[\sum_{j=1}^{N} U_j \cos(j\omega) \right]^2 \right\}^{1/2}, \quad (3)$$

where ω is the angle between adjacent side chains when the sequence is considered as a regular structure and viewed down an axis defined by the Cα atoms. If the val-

Table 1. *Difference matrix calculated from the substitution tables for buried residues in water-soluble proteins and for the lipid-accessible residues* [a]

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	+.274	−.005	+.027	+.088	−.071	+.040	+.012	+.013	+.019	−.039	+.010	+.046	−.165	−.032	+.013	−.016	−.036	−.009	−.126	+.032
C	+.007	+.797	0	0	−.029	−.021	0	−.020	0	−.021	−.004	+.006	−.079	+.002	0	−.083	−.016	−.017	−.007	+.006
D	+.007	0	+.755	+.039	+.002	+.006	+.005	+.008	+.019	+.002	+.002	+.144	+.005	+.011	0	+.016	+.009	+.007	0	+.006
E	+.007	+.011	+.013	+.442	+.002	+.008	0	+.003	0	0	+.015	+.036	0	+.018	+.013	+.007	+.005	+.008	+.001	0
F	−.062	−.048	+.001	+.007	+.313	−.132	+.023	−.094	0	−.032	−.059	+.014	+.020	−.010	+.006	−.064	−.137	−.090	−.052	+.070
G	−.006	−.020	+.013	+.039	−.026	+.436	0	−.040	0	−.063	−.014	−.040	+.006	+.013	+.019	−.044	−.087	−.059	−.014	+.003
H	+.005	0	+.001	0	−.039	+.003	+.722	+.004	+.019	+.004	+.005	+.044	+.005	+.059	+.006	+.012	+.002	+.005	+.006	+.013
I	−.043	−.037	+.015	+.025	−.082	−.189	+.008	+.132	+.028	+.024	+.026	−.083	+.015	+.006	+.006	−.032	−.008	+.026	−.103	−.359
K	+.012	+.010	+.001	+.004	+.003	+.005	+.002	+.004	+.462	+.004	+.006	+.020	+.016	+.013	+.144	+.009	+.009	+.004	0	+.003
L	−.186	−.287	+.004	+.049	+.031	−.123	+.030	−.020	+.009	+.273	+.073	−.105	+.021	−.094	+.045	−.122	−.138	−.096	−.033	+.021
M	−.033	−.012	+.002	+.039	−.001	−.054	+.016	−.001	0	+.028	+.044	−.067	+.001	+.059	0	−.006	−.018	+.008	−.121	−.018
N	+.012	+.002	+.083	+.014	+.003	+.002	+.025	+.003	0	+.003	+.006	+.386	+.014	−.035	+.006	+.025	+.013	+.004	+.004	−.026
P	−.026	+.003	+.005	0	−.048	+.013	+.003	+.007	+.028	+.002	+.001	+.022	+.053	+.011	+.006	+.010	+.019	+.007	0	+.003
Q	+.014	+.010	+.007	+.092	+.002	+.010	+.036	+.002	+.057	−.034	+.032	+.018	+.001	−.035	+.104	+.003	+.003	+.007	+.005	+.004
R	+.009	+.009	0	+.042	+.005	+.004	+.012	+.004	+.311	+.003	+.002	0	+.003	+.032	+.612	+.010	+.003	+.004	+.002	+.005
S	−.054	−.243	+.017	+.032	−.034	+.038	+.018	−.034	+.009	−.057	−.021	−.010	+.020	+.017	0	+.359	+.073	−.001	−.013	+.007
T	+.001	−.100	+.013	+.004	−.079	+.003	+.008	−.060	+.019	−.074	−.147	−.162	+.040	−.079	+.006	+.015	+.315	−.050	−.017	−.157
V	+.071	−.094	+.026	+.067	+.028	−.042	+.046	+.099	0	+.040	+.020	−.301	+.016	+.027	+.013	−.093	−.016	+.278	−.097	−.029
W	−.011	+.002	+.005	+.004	−.036	−.009	+.008	−.019	0	−.066	−.015	+.004	0	+.004	0	−.005	−.006	−.049	+.605	−.057
Y	+.002	+.003	+.012	+.014	+.057	+.001	+.026	+.010	+.019	+.005	+.020	+.028	+.008	+.013	0	−.004	+.011	+.014	−.041	+.472

[a] Positive values represent substitutions more likely for buried residues.

ues of U_j represent the hydrophobicities (H_j) of the residues, then M is the hydrophobic moment (μ; Eisenberg et al., 1984).

When calculating the periodicity in the values of U_j, the Fourier transform power spectrum is calculated by

$$P(\omega) = [x]^2 + [y]^2, \tag{4}$$

where

$$x = \sum_{j=1}^{N} U_j^n \sin(j\omega), \tag{5}$$

$$y = \sum_{j=1}^{N} U_j^n \cos(j\omega), \tag{6}$$

and where

$$U_j^n = U_j - \bar{U} \quad (j = 1, 2, \ldots, N). \tag{7}$$

$\bar{U}$ is the average value of U_j over the window. U^n is therefore a normalized version of the profile U adjusted so that $\sum_{j=1}^{N} U_j^n = 0$ and hence $\bar{U}^n = 0$. The new profile U^n consists of elements U_j^n in which the periodicity in internal/external residues is predicted by the periodicity in positive/negative values (or negative/positive values) of U_j^n. This results in cleaner power spectra (since $P(\omega) = 0$ when $\omega = 0$) so that the alpha periodicity index AP can be calculated as

$$\mathrm{AP} = \frac{(1/30)\int_{90°}^{120°} P(\omega)\, d\omega}{(1/180)\int_{0°}^{180°} P(\omega)\, d\omega}. \tag{8}$$

AP is analogous to ψ used in Komiya et al. (1988) and to the amphipathic index AI used in Cornette et al. (1987) (although the precise boundaries of the helical regions of the power spectrum differ in the latter). AP is a ratio of the extent of the periodicity in the helical region of the spectrum compared with that over the whole spectrum. Komiya et al. (1988) suggest that a value of AP greater than 2 indicates that the helical periodicity is significant. This suggests the presence of a helix within the window of sequence used, but it is difficult to predict the precise start and finish of the helix.

The direction of the internal face of a predicted helix can be estimated from the direction of the moment $\sqrt{P(\omega)}$ when $\omega = 100°$. This is the moment produced by the profile U^n when the sequences form an ideal α-helix. θ is the angle describing the direction of the moment relative to the first residue ($j = 1$). θ can be calculated by

$$\gamma = \arccos[y/\sqrt{P(\omega)}], \tag{9}$$

where γ is greater than 0° and less than 180°.

$$\theta = \begin{cases} \gamma, & \text{if } x > 0; \\ 360 - \gamma, & \text{if } x < 0. \end{cases} \tag{10}$$

Modifications to the standard Fourier transform procedure

Although the detection of periodicity is increased by normalizing U to U^n (Equation 7), there are still potential problems that can distort the value of AP in certain situations. Because this may result in a value of AP below the limit of significance (2), we have included three modifications to the above method in order to reduce the occurrence of such distortions. The modifications result in a clearer detection of the periodicity in the profile U, be it helical or otherwise. This not only results in a more reliable detection of helical regions but also results in a more reliable value of θ because this value is only meaningful if the helical periodicity is significant.

Treatment of outliers

The first potential problem is caused by one or more unusually high or low value of U_j. For example, such a situation may occur when U_j represents the variability profile V_j (Komiya et al., 1988) and where one particular position is completely conserved (say for a functional purpose) in an otherwise highly divergent family of sequences. We refer to these high or low values of U_j as outliers and identify them as those values of U_j where $|\bar{U} - U_j| > |\bar{U} - 2\sigma|$ (σ = the standard deviation in the elements of U). These outliers need to be smoothed because they can distort the value of AP and so mask the underlying periodicity.

This situation is depicted in Figure 2A where a profile U is plotted with the average value $\bar{U} = 0$ (left). The power spectrum for this profile is shown on the right where AP = 2.15. The dashed lines parallel to the x axis represent the boundaries of 2σ, and the dashed line through the points represents the best least-squares fitted line. U_j at position $j = 4(U_4)$ is an outlier, which causes the average value $\bar{U}$ to be higher than it would otherwise be. The profile U is smoothed to give U^s in which the outlying values U_j are adjusted to U_j^s by

$$U_j^s = \begin{cases} U_{\max} & \text{if } U_j > \bar{U}; \\ U_{\min} & \text{if } U_j < \bar{U}, \end{cases} \tag{11}$$

where

$$U_{\max} = \max[U_j] \quad \text{for each } |\bar{U} - U_j| \leq |\bar{U} - 2\sigma| \tag{12}$$

and

$$U_{\min} = \min[U_j] \quad \text{for each } |\bar{U} - U_j| \leq |\bar{U} - 2\sigma|. \tag{13}$$

In the example, U_4 is given the value of U_1 because this is the highest value of U_j within the bounds of 2σ. Equations 4–8 can then be used with this smoothed profile U_j^s (Fig. 2B; left) to recalculate the power spectrum and AP (right). The periodicity about the average value is more visible with the smoothed values, and AP is increased to 3.41.

Treatment of ramps

A second improvement can be made to the procedure. It is evident from Figure 2B (left) that the best least-squares fitted line through the points U_j^s (dashed line) is not parallel to the x axis (i.e., it is ramped). The best least-squares fitted line through the elements represents the midpoint of U^s more accurately than the average value $\bar{U}^s$. Because we are actually calculating the periodicity about the average value (due to Equation 7), the elements of U^s are adjusted to give the profile U^{gs} for which the gradient of the best least-squares fitted line is zero (Fig. 2C; left). The average value and the best least-squares fitted line are therefore equivalent. The profile U^{gs} is used in the same way as U to calculate a new power spectrum (Fig. 2C; right) using Equations 4–8. This procedure removes trends in the data that may produce spurious low frequency peaks in the power spectrum that can decrease AP and so may mask the helical periodicity. The profile U^{gs} has an improved value of AP = 3.71.

Therefore in our method, U (for each window of the sequence alignment) is first smoothed to U^s, which is then adjusted to give U^{gs}. The periodicity in this profile can then be calculated using Equations 4–8. Equation 7 converts the profile U^{gs} to U^{ngs}, which has an average value of zero. The elements U_j^{ngs} represent the extent to which the residue at position j is likely to be buried and it is the periodicity in these values that we use to predict helices.

Modification to variability method

There is a possibility that a proline may be conserved on the lipid-facing side of a transmembrane helix (possibly to maintain a bend; Barlow & Thornton, 1988) and hence the Fourier transform of Komiya et al. (1988) will be disrupted because it assumes conserved residues are buried. Therefore, if a position in a sequence alignment contains a *conserved* proline, we use a slightly modified version of the method described in Komiya et al. (1988) by adjusting the profile V so that

$$\bar{V} = \bar{V}^{-\text{PRO}}, \tag{14}$$

where $\bar{V}^{-\text{PRO}}$ is the average value of V_j for all positions in the window excluding that of the conserved proline. Using this new value of $\bar{V}$,

$$V_j^{\text{PRO}} = \bar{V}, \tag{15}$$

where V_j^{PRO} is the new value of V_j at the position of the conserved proline. Thus a conserved proline does not bias $P(\omega)$ because $V_j^{\text{PRO}} - \bar{V} = 0$. This procedure is also carried out when C_j ($C_j = 21 - V_j$) is being used.

In many cases these three modifications to the standard Fourier transform method will have little or no effect on the value of AP because there may be no outliers or conserved prolines, and the gradient of the profile may be close to zero. However, in the cases when any of these situations result in a low value of AP, the method is still able to detect the underlying periodicity. It should be noted that the modifications do not necessarily boost the value of AP, but rather they make the periodicity in U clearer, whether or not this is typical of a helix.

Application of the Fourier transform method

We use the substitution tables to predict whether a position in a sequence alignment is buried or exposed. Using the difference matrix D^{bl} described above (Table 1), the difference probability profile S (composed of elements S_j) is calculated. S_j is the average value of the difference probabilities for all pairwise substitutions at each position j in the alignment. Therefore, for an alignment of y sequences, there are $y(y-1)$ pairwise substitutions possible. However, sequences that are identical over the window being used are included only once.

For a window of length N over the sequence alignment, the periodicity in the values of S_j can be calculated using the Fourier transform procedure described above. We use a window range of 7–12 residues to test the method on a sequence alignment of bacteriorhodopsin and two related proteins. The values of AP and ψ quoted for each helix are the best values obtained within this window range. This optimal window (W_{max}) is used to calculate θ, and the entire transmembrane helix is constructed based upon the phase of this region. Once the lipid-facing side of the helix has been identified, the point at which this face contacts the more polar environment of the phosphate head groups and surrounding aqueous solution can be estimated by the point at which charged residues appear on this side of the helix. Charged residues should not be present on the lipid-facing side of a transmembrane helix.

The results of a prediction can be tested by comparing the value of θ calculated from the profile S (or H, V, C) in the prediction, with θ calculated from the percentage side-chain accessibilities (a) computed from the known structure. A profile I is calculated from the structure and consists of elements I_j where $I_j = -a$ for the residue at position j. The direction of the moment calculated from I indicates the internal side of the helix using information derived from the structure of the protein.

We have developed a suite of FORTRAN programs, PERSCAN, that can be used to search for helical periodicity in sequence alignments and predict the internal face of any helix found. Substitution tables (S_j), hydrophobicity scales (H_j), variability (V_j), conservation (C_j) or accessibility profiles (I_j) can be used in the programs so that the results of each method can be compared.

Results and discussion

Lipid-accessible substitution tables

The substitution patterns for each type of amino acid when buried (water-soluble proteins), aqueous-accessible, and lipid-accessible are shown in Figure 3. The patterns for lipid-accessible residues are based on 3,853 pairwise substitutions and show that they are less conserved than the equivalent buried residues in water-soluble proteins, as indicated by the predominantly positive values on the diagonal in Table 1. There are no charged residues (Asp, Glu, His, Lys, Arg) on the lipid-facing side of the 22 helices of the reaction center structures used in this analysis. Moreover, the sequence alignments show that uncharged residues that face the lipid in one of the structures are very rarely substituted by charged residues in related sequences.

The statistical significance of the results for the lipid-facing residues depends on the residue type as indicated by the error bars in Figure 3. The results for Cys, Asn, Pro, Gln, and Tyr are the least significant because they rarely occur on the lipid-facing side of the helices in the structures available. The more abundant hydrophobic residues (Phe, Ile, Leu, Val, Trp) give more reliable substitution patterns. As more structures of membrane proteins become available, the tables will become more general and hence more significant.

The substitution tables show distinct patterns depending on the environment. For example, buried isoleucine residues are usually conserved, but when they are substituted, it is usually by valine or leucine. Solvent-accessible isoleucine residues are less conserved, and although valine and leucine are still the most probable substitutions, there is a general increase in the probability of substitutions to nearly all the other amino acid types. In the case of lipid-accessible isoleucine residues, there is still a reduction in the conservation but there is not a general increase in substitutions to other amino acids but rather an increase only in substitutions by uncharged residues. This reduction in the conservation and the absence of substitutions by charged residues is typical of the substitution patterns for lipid-facing residues. Therefore, a conserved charged residue generally has a greater effect on θ than a conserved uncharged residue. This adds an extra dimension to the method of Komiya et al. (1988), which only considers variability and assumes that each residue type is equivalent.

Prediction of bacteriorhodopsin helices

The structure of bacteriorhodopsin has been defined using high-resolution electron cryomicroscopy (Henderson et al., 1990). Bacteriorhodopsin has seven transmembrane helices, each with one side buried and the other side exposed to the lipid, and hence it provides an appropriate test for

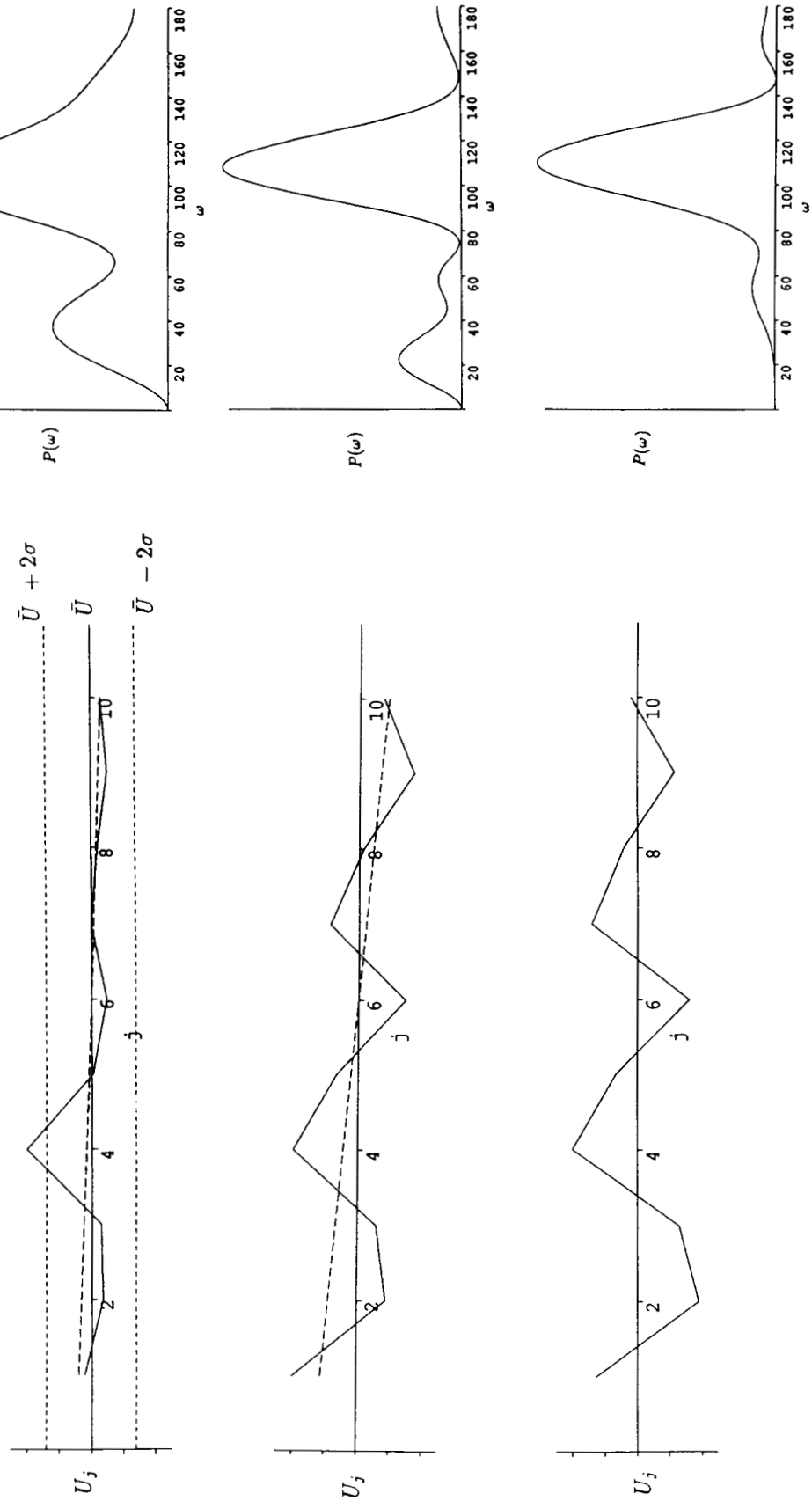

Fig. 2. **A:** An example of a profile U over a 10-residue window where the average $\bar{U}$ is zero (left). The dashed lines parallel to the x axis are the boundaries of 2σ. U_4 is an outlier that masks the periodicity. The power spectrum (right) for this profile has AP = 2.15. **B:** Smoothing U_4 to the value of U_1 (see text) improves the periodicity and the power spectrum is cleaner (right) with AP = 3.41. The best least-squares fitted line through the elements U^s is shown by the dashed line (left). **C:** Profile U^{gs} (see text) with the best least-squares fitted line equivalent to the x axis (left) shows an improved power spectrum (right) with a reduction in the low frequency peaks. AP = 3.71.

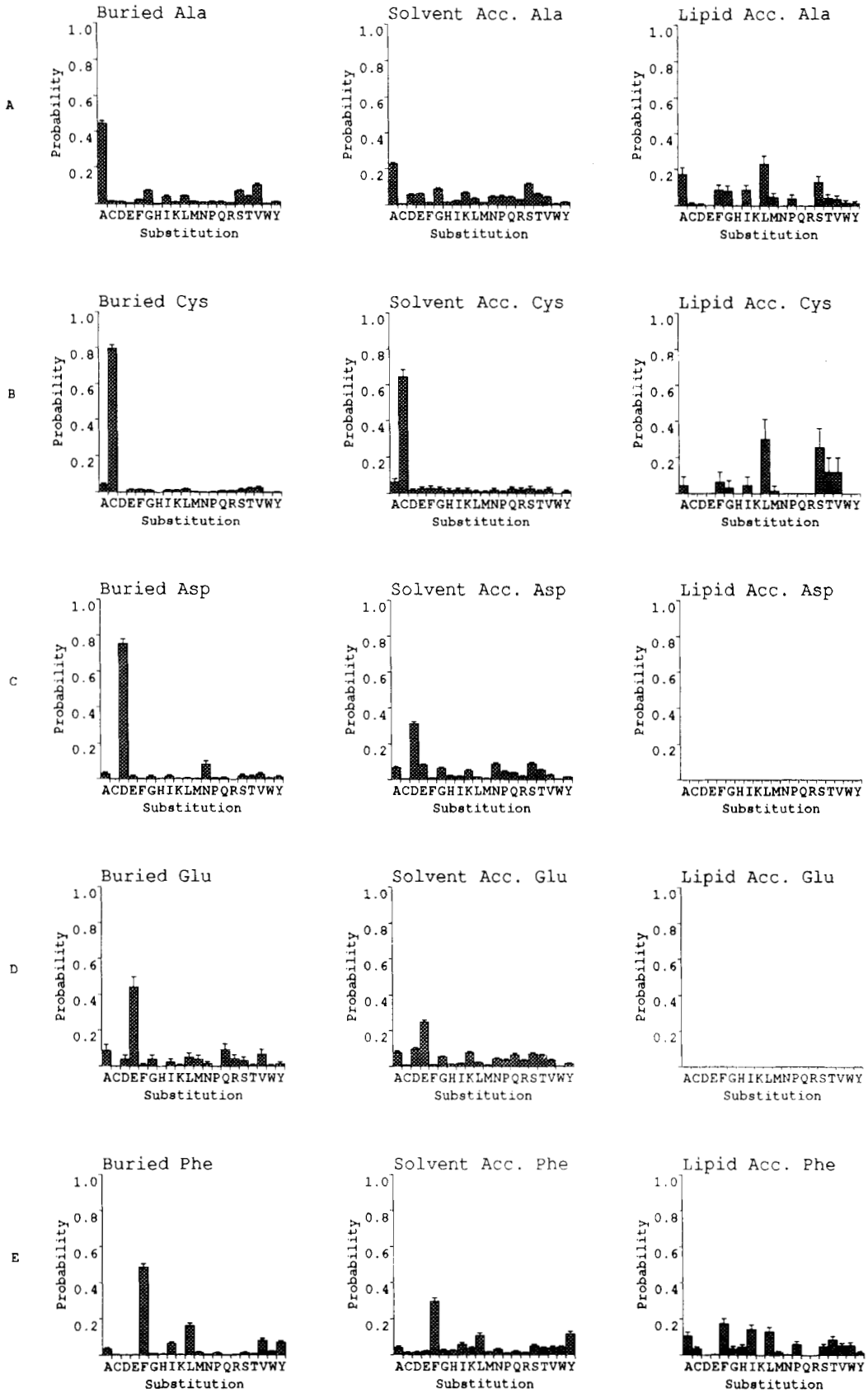

Fig. 3. A–T: Individual substitution tables for buried, solvent (water)-accessible, and lipid-accessible residues. Buried residues tend to be more conserved than both solvent-accessible and lipid-accessible residues. Lipid-accessible residues tend to rarely substitute to charged residues. (*Figure continues on next three pages.*)

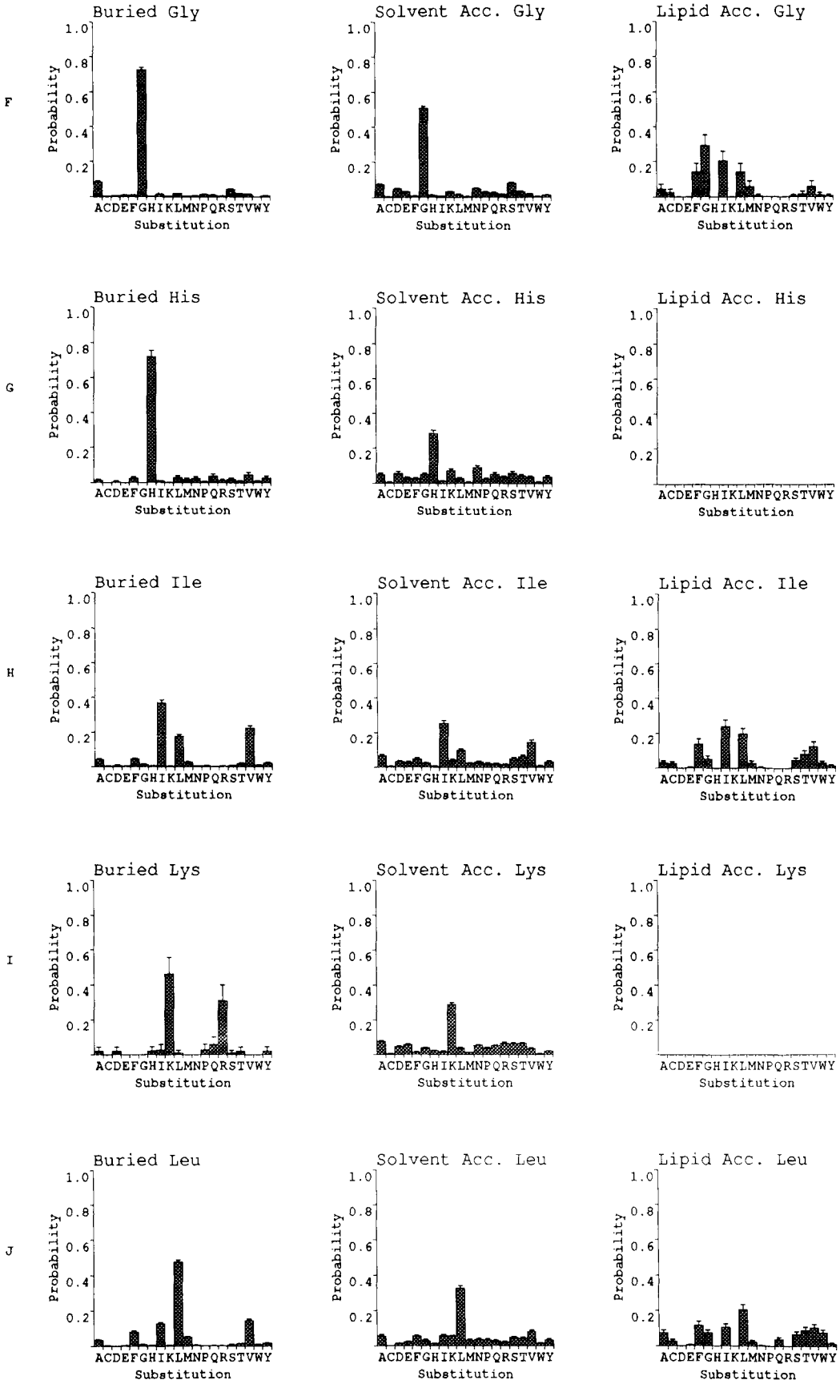

Fig. 3. *Continued.*

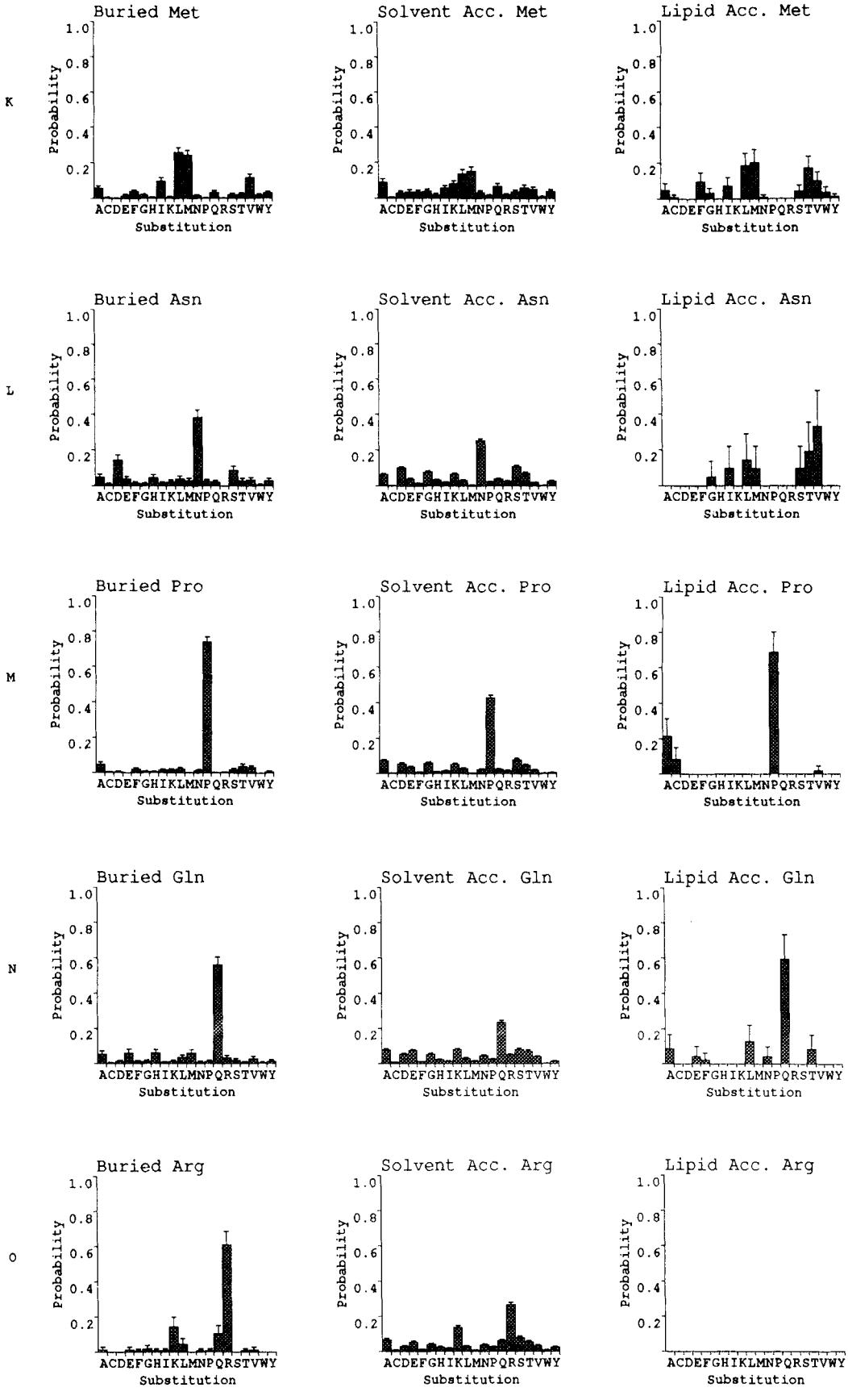

Fig. 3. *Continued.*

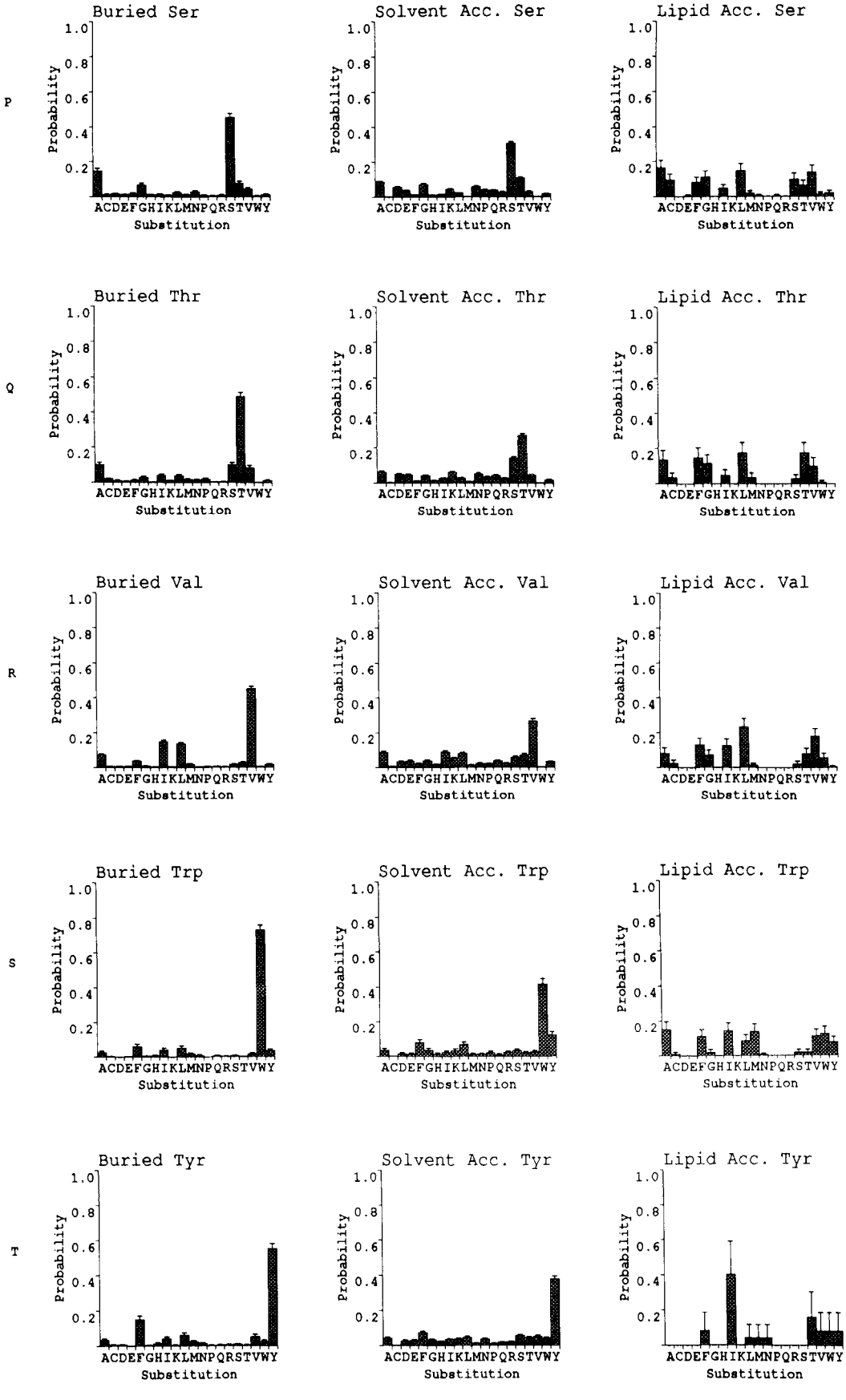

Fig. 3. *Continued.*

the Fourier transform procedure described above. Bacteriorhodopsin is homologous to two other proteins found in halobacteria: halorhodopsin and sensory rhodopsin. An alignment of these three protein sequences can be found in Henderson et al. (1990).

The alignment of the seven transmembrane helices was used in the test of the prediction method outlined above, and the results are summarized in Tables 2 and 3. The Fourier transform method used in Komiya et al. (1988) (i.e., without the three modifications described above), using a single scanning window size of 18, detected only four of the seven helices. However, the use of a variable window size (7–12) and the three modifications results in the detection of all seven helices. This is also the case when profile S is used.

A comparison of the direction of the moment calculated from the substitution tables (S) with that calculated from the side-chain accessibilities (I) shows that the internal side of each helix is correctly predicted. Again, the results are similar to those obtained using variability (V_j). The worst result obtained was for helix 5, where the difference in the directions of the moments was 52° using S_j and 39° using V_j.

Figure 4 shows data for helix 7 in more detail. The alignment of the three sequences is shown in Figure 4A, and the optimal window is indicated by the dashed line. The Fourier transform calculated from I (Fig. 4B; left) shows a strong peak at about 100° with AP = 4.10, and this is similar to the Fourier transform calculated using S, which has a value of AP = 3.42 (Fig. 4B; right). The values of I_j^{ngs} and S_j^{ngs} can be plotted for each residue as vectors in the form of a helical wheel with 3.6 residues per turn. The vector wheels for helix 7 are shown in Figure 4C for I_j^{ngs} (left) and for S_j^{ngs} (right). The vector for the first residue in the window is plotted at 12 o'clock and the remaining vectors are plotted clockwise at 100° intervals. The sum of these vectors when $\omega = 100$ gives the moment $\sqrt{P(\omega)}$. The helical wheels (Fig. 4D) indicate the direction of this moment, which represents the internal face of the helix. The difference between the direction of the moment calculated from the side-chain accessibility (profile I; left) and that from the prediction (profile S; right) provides a crude measure of the accuracy of the prediction. In this case the moments differ by less than 2°, and the chromophore binding lysine (K-216) is clearly positioned on the inside face of the helix.

The helical wheels can be plotted in a vertical fashion as shown in Figure 4E so that the horizontal lines projecting from the sequence represent the extent to which that position is buried in the core of the protein. Horizontal lines to the right indicate buried positions and lines to the left indicate exposed positions. The window that results in the optimal value of AP is indicated by solid lines and capital letters. The asterisks (*) represent positions where there is at least one charged residue in the alignment, and these should be absent from the lipid-facing side of the helix unless they are able to contact the phosphate head group region of the bilayer or the aqueous environment outside the membrane. Figure 4E shows the vertical helical plot for helix 7 calculated from profile I (left) and profile S (right), and it can be seen that although there are charged residues on the buried face of the helix, there is an absence of charged residues over a region of 21 residues on the lipid-facing side. This charge-free region can be used to identify the membrane-buried portion of transmembrane helices and the points at which they reach the more polar regions at either side of the membrane.

The orientation of each of the seven helices in bacteriorhodopsin was predicted and because in this case the topology of the protein is known, the predictions were used to place the individual helices into an approximate representation of the protein fold. Figure 5A shows the seven predicted helices arranged in an antiparallel fash-

Table 2. *Alpha periodicity (AP) indices* [a]

Helix	(i)	(ii)	(iii)
1	1.02	3.04	2.52
2	2.37	3.88	3.21
3	3.50	3.79	3.72
4	3.88	3.89	3.65
5	1.60	3.28	3.72
6	2.59	4.32	3.77
7	1.24	2.90	3.42

[a] AP (Equation 8, see text) calculated from the alignment of each of the seven helical regions using (i) the profile V, a fixed scanning window size of 18, and the standard Fourier transform approach of Komiya et al. (1988); (ii) the profile V, a variable scanning window size of 7–12, and the three modifications described in the text; (iii) the profile S, a variable scanning window size of 7–12, and the three modifications described in the text. The modified method results in the detection of all seven helices, whereas the original approach detects only four. The values calculated using the substitution tables (S_j) are comparable with those calculated from variability (V_j).

Table 3. *Difference (in degrees) between the direction of the internal face of the predicted helices (from profiles V and S) and that calculated from the structure (profile I)* [a]

Helix	V	S
1	16	14
2	26	42
3	11	30
4	16	9
5	39	52
6	13	5
7	13	2

[a] The correct face of each helix is predicted, and the two prediction methods give comparable results.

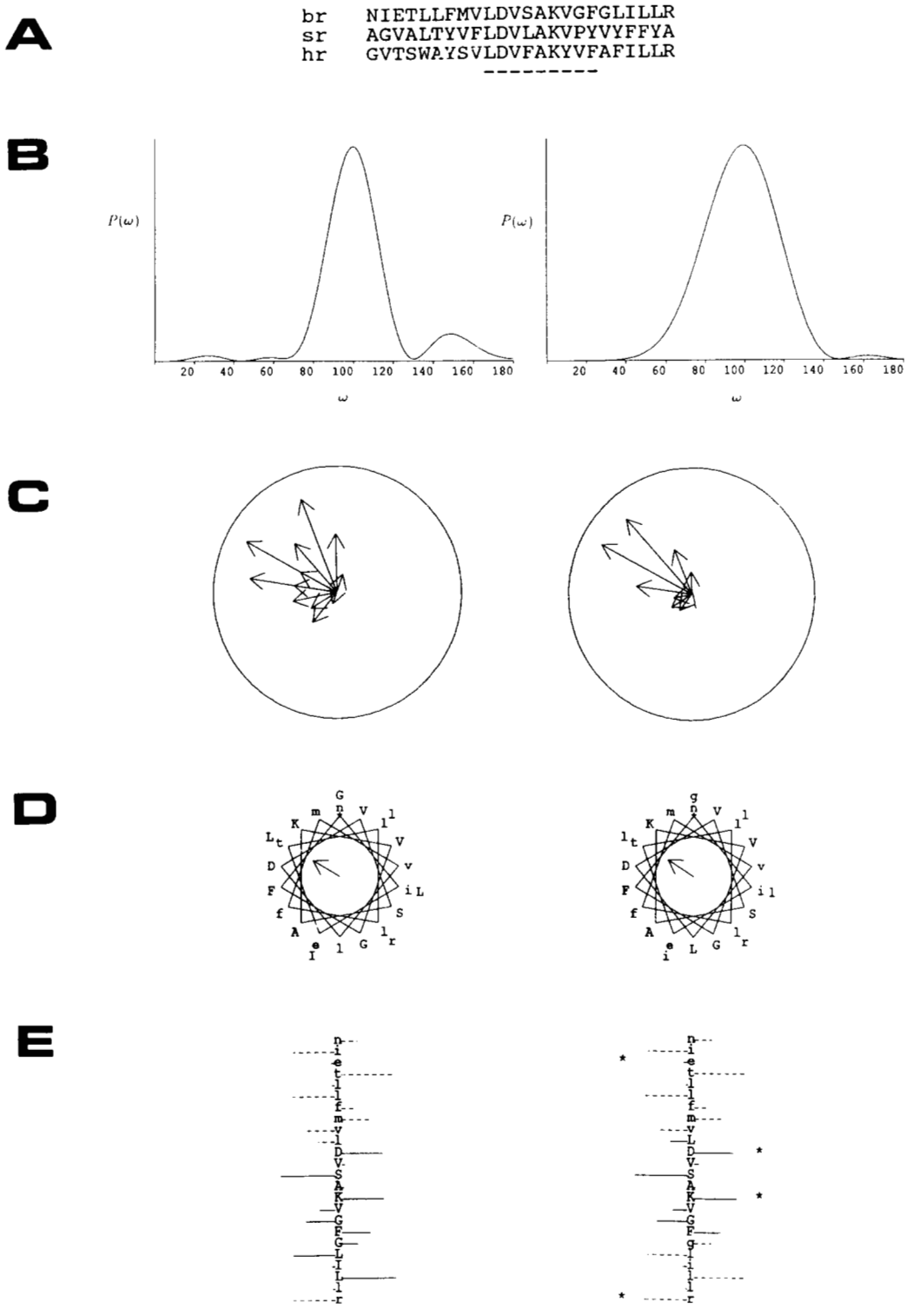

Fig. 4. A: Alignment of helix 7 from bacteriorhodopsin (br) with the equivalent regions in halorhodopsin (hr) and sensory rhodopsin (sr). The dashed line indicates the optimal window (range = 7–12) that produced the highest value of AP in the prediction. **B:** Optimal power spectrum calculated from I_j (left) and S_j (right) using a range of 7–12 for the window size. The prominent peak near 100° in both spectra indicates that this region is helical. **C:** Individual vectors I_j^{ngs} plotted as a helical wheel (left). The vectors point to the inside face of the helix, and this is compatible with that calculated from S_j^{ngs} (right). **D:** Helical wheel of helix 7. Residues within the optimal window are shown in uppercase. The arrow corresponds to the direction of the sum of the individual I_j^{ngs} (left) and S_j^{ngs} (right) vectors and represents the predicted internal face of the helix. **E:** Vertical representations of the helical wheels in Figure 4D. The N-terminal end of the helix is at the top, and the horizontal lines represent the extent to which that residue is buried or exposed. Lines to the right indicate buried residues, whereas lines to the left represent exposed residues. The residues in the optimal window (W_{max}) are in uppercase with solid horizontal lines.

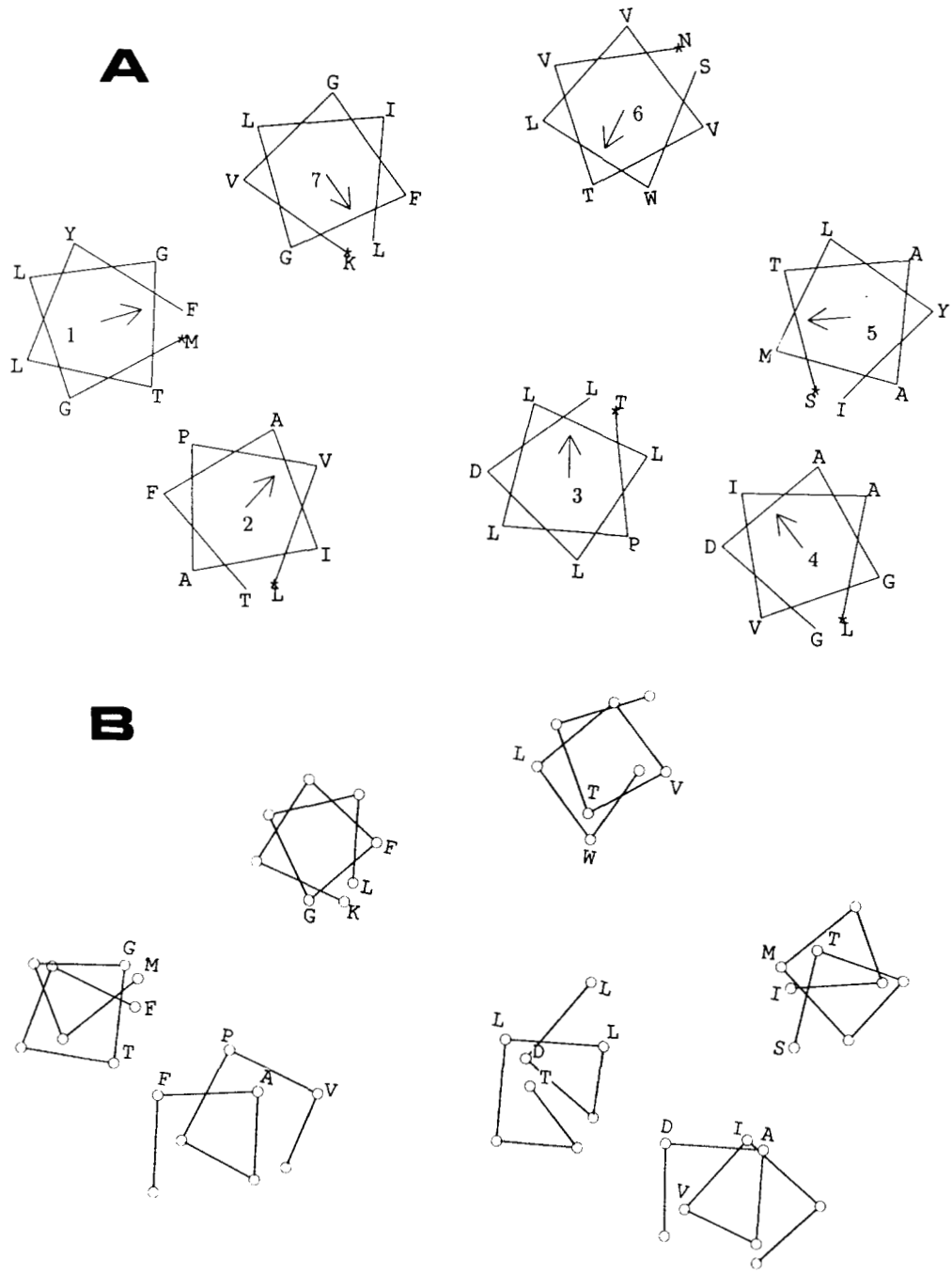

Fig. 5. A: The seven helices individually arranged in an antiparallel fashion with the predicted internal faces of each helix pointing to the interior of the bundle. The internal faces are compatible with those in the structure shown in **B**. Only eight residues are shown for clarity.

ion with the predicted inside face of each helix on the interior of the bundle. Only eight residues are shown for clarity, but the position of the remaining residues can be found by extrapolation. Figure 5B and Kinemage 1 show the equivalent residues in the bacteriorhodopsin structure. It can be seen that the prediction of buried and exposed residues from the sequence alignment correlates well with the structure.

Conclusions

Because only one protein family of transmembrane helices was available to generate the substitution tables for lipid-facing residues, we have made the assumption that the substitution pattern for residues in the core of membrane proteins will be similar to those of the buried residues in water-soluble proteins. This is necessary because

the substitution tables for buried residues from the reaction center will be biased by the constraints unique to its fold and function. For example, many residues in the reaction center core are important in packing the cofactors into a specific orientation necessary for electron transport across the membrane. The lipid-facing residues are less likely to be biased in this way because they are less important to the specific packing and function. Their primary role, to solubilize the protein in the lipid bilayer, is likely to be similar in other membrane proteins.

It is difficult to assess fully the accuracy of the method because there are so few membrane proteins with known structures. Bacteriorhodopsin provides the only suitable test case. Although the side-chain accessibilities are likely to be inaccurate (because the present structure is determined at only medium resolution), they are probably correct in general and are suitable for use in determining the internal face of each helix using the profile *I*. Only three sequences are available for this protein family, which also adds to the difficulty in using substitution tables to predict periodicity. However, despite these problems, the predictions show good agreement with the available structure and with a related prediction method using variability. The modifications to the Fourier transform method provide an increased chance of detecting the underlying periodicity and these can be applied to other related methods.

Because there are now many sequences available for membrane proteins, this method provides a useful way of predicting the position and orientation of helices in transmembrane regions. If the inside face of each helix is predicted and the depth to which the helices are buried in the membrane with respect to each other is known, then it is possible to construct useful three-dimensional models of the helical arrangement. This approach has been used to carry out a detailed modeling study on the family of G protein-coupled receptors and of bacterial light-harvesting proteins that will be published elsewhere (D. Donnelly, unpubl.).

Acknowledgments

We thank Prof. R. Henderson for providing us with the coordinates of bacteriorhodopsin and Prof G. Feher, Dr. J.P. Allen, and Prof D.C. Rees for sending us the coordinates of the reaction center from *R. sphaeroides*. We also thank various members of the Department of Crystallography, Birkbeck College, for their interesting comments and suggestions and also the S.E.R.C., the I.C.R.F, Pfizer, and Merck, Sharp and Dohme for their support.

References

Allen, J.P., Feher, G., Yeates, T.O., Komiya, H., & Rees, D.C. (1987a). Structure of the reaction center from *Rhodobacter sphaeroides* R-26: The cofactors. *Proc. Natl. Acad. Sci. USA 84*, 5730–5734.

Allen, J.P., Feher, G., Yeates, T.O., Komiya, H., & Rees, D.C. (1987b). Structure of the reaction center from *Rhodobacter sphaeroides*, R-26: The protein subunits. *Proc. Natl. Acad. Sci. USA 84*, 6162–6166.

Allen, J.P., Feher, G., Yeates, T.O., Komiya, H., & Rees, D.C. (1988). Structure of the reaction center from *Rhodobacter sphaeroides* R-26: Protein-cofactor (quinone and Fe^{2+}) interactions. *Proc. Natl. Acad. Sci. USA 85*, 8487–8491.

Barlow, D.J. & Thornton, J.M. (1988). Helix geometry in proteins. *J. Mol. Biol. 201*, 601–619.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Schimanovichi, T., & Tasumi, M. (1977). The Protein Data Bank: A computer based archival file for macromolecular structures. *J. Mol. Biol. 112*, 535–542.

Bleasby, A.J. & Wootton, J.C. (1990). Construction of validated, non-redundant composite protein sequence databases. *Protein Eng. 3*, 153–159.

Bowie, J.U., Reidhaar-Olson, J.F., Lim, W.A., & Sauer, R.T. (1990). Deciphering the message in protein sequences—Tolerance to amino acid substitutions. *Science 247*, 1306–1310.

Chang, C.-H., Tiede, D., Tang, J., Smith, U., Norris, J., & Schiffer, M. (1986). Structure of *Rhodopseudomonas sphaeroides* R-26 reaction center. *FEBS Lett. 205*, 82–86.

Chothia, C. & Lesk, A.M. (1986). The relationship between the divergence of sequence and structure in proteins. *EMBO J. 5*, 823–826.

Cornette, J.L., Cease, K.B., Margalit, H., Spouge, J.L., Berzofsky, J.A., & DeLisi, C. (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol. 195*, 659–685.

Deisenhofer, J., Epp, O., Miki, K., Huber, R., & Michel, H. (1984). Structure of the protein subunits in the photosynthetic reaction centre of *Rhodopseudomonas viridis* at 3 Å resolution. *J. Mol. Biol. 180*, 385–398.

Deisenhofer, J., Epp, O., Miki, K., Huber, R., & Michel, H. (1985). X-ray structure-analysis of a membrane-protein complex—Electron density map at 3 Å resolution and a model of the chromophore of the photosynthetic reaction centre from *Rhodopseudomonas viridis*. *Nature 318*, 618–624.

Donnelly, D., Johnson, M.S., Blundell, T.L., & Saunders, J. (1989). An analysis of the periodicity of conserved residues in sequence alignments of G protein-coupled receptors—Implications for the 3 dimensional structure. *FEBS Lett. 251*, 109–116.

Eisenberg, D., Weiss, R.M., & Terwilliger, T.C. (1984). The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. USA 81*, 140–144.

Engelman, D.M., Henderson, R., McLachlan, A.D., & Wallace, B.A. (1980). Path of the polypeptide chain in bacteriorhodopsin. *Proc. Natl. Acad. Sci. USA 77*, 2023–2027.

Engelman, D.M., Steitz, T.A., & Goldman, A. (1986). Identifying nonpolar transbilayer helices in amino-acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem. 15*, 321–353.

Henderson, R., Baldwin, J.M., Ceska, T.A., Zemlin, F., Beckmann, E., & Downing, K.H. (1990). Model for the structure of bacteriorhodopsin based on high-resolution electron cryomicroscopy. *J. Mol. Biol. 213*, 899–929.

Hubbard, T.J.P. & Blundell, T.L. (1987). Comparison of solvent-inaccessible cores of homologous proteins—Definitions useful for protein modelling. *Protein Eng. 1*, 159–171.

Jennings, M.J. (1989). Topography of membrane proteins. *Annu. Rev. Biochem. 58*, 999–1027.

Komiya, H., Yeates, T.O., Rees, D.C., Allen, J.P., & Feher, G. (1988). Structure of the reaction center from *Rhodobacter sphaeroides* R-26: Symmetry relations and sequence comparisons between different species. *Proc. Natl. Acad. Sci. USA 85*, 9012–9016.

Kühlbrandt, W. & Wang, D.N. (1991). Three-dimensional structure of plant light-harvesting complex determined by electron crystallography. *Nature 350*, 130–134.

Lee, B. & Richards, F.M. (1971). The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol. 55*, 379–400.

Overington, J., Donnelly, D., Johnson, M.J., Šali, A., & Blundell, T.L. (1992). Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. *Protein Sci. 1*, 216–226.

Overington, J.P., Johnson, M.S., Šali, A., & Blundell, T.L. (1990). Tertiary structural constraints on evolutionary diversity: Templates, key residues and structure prediction. *Proc. R. Soc. Lond. B 241*, 132–145.

Rees, D.C., DeAntonio, L., & Eisenberg, D. (1989a). Hydrophobic organization of membrane proteins. *Science 245*, 510–513.

Rees, D.C., Komiya, H., Yeates, T.O., Allen, J.P., & Feher, G. (1989b). The bacterial photosynthetic reaction center as a model for membrane-proteins. *Annu. Rev. Biochem. 58*, 607–633.

Šali, A. & Blundell, T.L. (1990). Definition of general topological equivalence in protein structures: A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol. 212*, 403–428.

Smith, E.L. (1968). The evolution of proteins. *Harvey Lect. 62*, 231–256.

Wallace, B.A. (1990). Gramicidin channels & pores. *Annu. Rev. Biophys. Biophys. Chem. 19*, 127–157.

Weiss, M.S., Kreusch, A., Schiltz, E., Nestel, U., Welte, W., Weckesser, J., & Schulz, G.E. (1991). The structure of *Rhodobacter capsulatus* at 1.8 Å resolution. *FEBS Lett. 280*, 379–382.

Yeates, T.O., Komiya, H., Chirno, A., Rees, D.C., Allen, J.P., & Feher, G. (1988). Structure of the reaction center from *Rhodobacter sphaeroides* R-26: Protein-cofactor (bacteriochlorophyll, bacteriopheophytin, and carotenoid) interactions. *Proc. Natl. Acad. Sci. USA 85*, 7993–7997.

Yeates, T.O., Komiya, H., Rees, D.C., Allen, J.P., & Feher, G. (1987). Structure of the reaction center from *Rhodobacter sphaeroides* R-26: Membrane-protein interactions. *Proc. Natl. Acad. Sci. USA 84*, 6438–6442.