PROTEINS: Structure, Function, and Genetics 18:281–294 (1994)

A Method for α-Helical Integral Membrane Protein Fold Prediction

William R. Taylor, David T. Jones, and N. Michael Green
Laboratory of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, U.K.

ABSTRACT **Integral membrane proteins (of the α-helical class) are of central importance in a wide variety of vital cellular functions. Despite considerable effort on methods to predict the location of the helices, little attention has been directed toward developing an automatic method to pack the helices together. In principle, the prediction of membrane proteins should be easier than the prediction of globular proteins: there is only one type of secondary structure and all helices pack with a common alignment across the membrane. This allows all possible structures to be represented on a simple lattice and exhaustively enumerated. Prediction success lies not in generating many possible folds but in recognizing which corresponds to the native. Our evaluation of each fold is based on how well the exposed surface predicted from a multiple sequence alignment fits its allocated position. Just as exposure to solvent in globular proteins can be predicted from sequence variation, so exposure to lipid can be recognized by variable-hydrophobic (*variphobic*) positions. Application to both bacteriorhodopsin and the eukaryotic rhodopsin/opsin families revealed that the angular size of the lipid-exposed faces must be predicted accurately to allow selection of the correct fold. With the inherent uncertainties in helix prediction and parameter choice, this accuracy could not be guaranteed but the correct fold was typically found in the top six candidates. Our method provides the first completely automatic method that can proceed from a scan of the protein sequence databanks to a predicted three-dimensional structure with no intervention required from the investigator. Within the limited domain of the seven helix bundle proteins, a good chance can be given of selecting the correct structure. However, the limited number of sequences available with a corresponding known structure makes further characterization of the method difficult.**

Key words: membrane, protein, structure, prediction, G-protein coupled receptor, rhodopsin

INTRODUCTION

The compartmentalization of cellular and subcellular spaces by phospholipid bilayer membranes is a fundamental feature of the organizational structure of living processes. However, little would be achieved by this strategy if communication and selective transport were not possible across the membranes. This traffic is almost entirely controlled by proteins about which very little is known structurally. The few structures that are known for integral (or multiple membrane spanning) proteins suggest a basic form composed of bundles of α-helices that span the membrane and pack together with their axes approximately normal to the plane of the membrane.[17]* Fortunately, the required length and the hydrophobic nature of these helices make their detection from sequence data a relatively simple matter—compared to, say, secondary structure prediction.[20,28] Unfortunately, a single sequence provides virtually no further guide as to how the helices should be packed relative to one another and any further information must be obtained by other methods.[27]

Many attempts have been made to model membrane proteins based on sequence similarity to one of the known structures (commonly bacteriorhodopsin[16]).[5,8,10,11] Other attempts have put forward plausible suggestions[14,15] for larger assemblies but these did not result form an automated procedure, making the uniqueness of their solution difficult to evaluate.

More recently, Baldwin[3] has proposed a model for the structure of rhodopsin. However, unlike the pre-

*The term *membrane protein* is used to describe this class of structure, ignoring the structure of porin which comprises a 16-strand β-barrel.

Received July 30, 1993; revision accepted November 4, 1993.
Address reprint requests to Dr. William R. Taylor, Laboratory of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, U.K.
D.T. Jones is jointly at Biomolecular Structure Unit, Department of Biochemistry, University College London, Gower St., London WC1E 6BT, U.K.

ceding studies, she avoided the common assumption that the G protein coupled receptor superfamily has the same fold as bacteriorhodopsin. This supposition is based only on the predicted number of helices and the binding of a light sensitive chromophore (retinal) in the opsins and bacteriorhodopsin. Both these correspondences may be coincidental especially as the retinal conformations are distinct in the two systems and one protein is a proton-pump while the other is a signal transducer. There is no explicit sequence similarity between the two families and any little similarity that can be imaged does not appear to support their direct correspondence.[22] Baldwin[3] used a semirigorous, partially automated method based on the analysis of conservation in multiple aligned sequences to arrive at a preferred structure that can claim to be a prediction as distinct from a "homology model."

Compared to the prediction of the tertiary structure of globular (water-soluble) proteins the prediction of the structure of membrane proteins is, in principle, a much easier problem. The constraint of longitudinal[†] helix packing reduces the solution-space of the problem from three dimensions to two. Despite the relative simplicity of the two-dimensional world and the central importance of the proteins involved, relatively little attention has been focused on prediction of membrane protein structure. This neglect may have resulted through of a lack of interest in the limited variety of anticipated forms, or (more probably) because the empirical based prediction methods currently used for globular proteins require a large data base from which to derive their rules and only a few membrane protein structures are known.

In this investigation, we direct some programs originally developed for the prediction of globular protein structure[25,26] toward the prediction of membrane protein structure. The adaptation of the methods has been guided by the work of Baldwin[3] on the G-protein coupled receptor superfamily and were applied initially, to bacteriorhodopsin (for which a structure is known) and then to the rhodopsin/opsin family which are G-protein coupled receptors.

Besides the fundamental difference in overall architecture[‡] of membrane protein compared to globular proteins, the essential difference in reapplication of the earlier methods has been the substitution of variable hydrophilic positions in a multiple sequence alignment (as an indicator of exposure to solvent) with variable hydrophobic positions (as indicative of exposure to lipid). Otherwise, the methods remain the same involving the combinatoric enumeration of windings over a predefined architecture, followed by the selection of a few preferred folds and their final elaboration into a realistic (α-carbon) model.[25,26]

METHODS

Sequence Analysis

Data selection and alignment

Six bacteriorhodopsin sequences were extracted from the SWISS-PROT sequence databank[2] by keyword extraction using the Wisconsin GCG software package.[9] These were aligned using the multiple sequence alignment program **MULTAL**[24] using an amino acid relatedness matrix derived from a large number of transmembrane segments from proteins with more than one membrane spanning helix[18,21] (Fig. 1).

G-protein coupled receptor sequences have no common "key-word" and were extracted by regular expression matching across the SWISS-PROT databank-using the AWK language as implemented in the **GAWK** program[4] with the pattern defined in the PROSITE databank.[1] These two were aligned by **MULTAL**. The resulting alignment of 30 sequences corresponded to that summarized by Baldwin[3] although being much smaller in size.

The prediction of transmembrane segments allowed further constraints to be applied to the alignments. Two residues defined as being in a transmembrane helix were given a slight bias to preferentially align and any gaps in these regions were discouraged. These constraints made little difference to most of the alignments but eliminated a few poorly justified insertions at the ends of some helices.

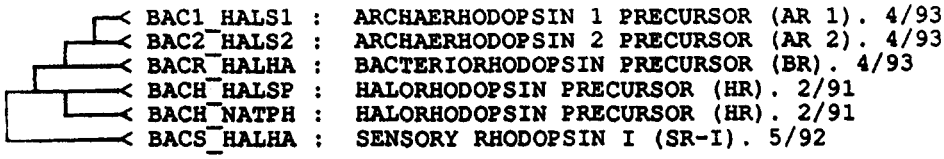
Hydrophobicity and conservation

The fundamental distinction between this application and equivalent methods developed for globular proteins is the definition of a measure of surface exposure. As previously, a simple average hydrophobicity was calculated for each position in a multiple alignment. However, the many hydrophobicity scales derived from globular proteins were all avoided in favor of a measure of the preference of amino acids to be found in the middle section of single membrane spanning helices[20] since these must be lipid-exposed (unless they form part of a complex) (see Table I).

The conservation of residues at a position in a multiple alignment was similarly defined (as previously) from the pairwise sum of amino acid similarity but using a relatedness table derived from the helices of multiple spanning integral membrane proteins.[18,21] In her analysis of sequence variation, Baldwin[3] attaches importance to the differences between otherwise closely similar sequences and an additional feature was introduced into our conser-

[†]The term *longitudinal* is used to describe both the parallel and antiparallel packing of helices.

[‡]The term *architecture* is used to refer to the relative juxtaposition of secondary structure units independently of how they are connected.




```
          mdpialtaavgadllgd grpe   tLWLGIGTLLMLIGTFYFIVkgwgvtdkeareYYSITILVPGIASA
          mdpialqa  gfdllnd grpe   tLWLGIGTLLMLIGTFYFIArgwgvtdkeareYYAITILVPGIASA
          mlellptavegvsqaqitgrpe   WIWLALGTALMGLGTLYFLVkgmgvsdpdakkFYAITTLVPAIAFT
     msitsvpgvvdagvlgaqsaaavrenALLSSSLWVNVALAGIAILVFVYmgrtirpgrprLIWGATLMIPLVSIS
mtetlppvtesavalqaevtqrelfefvlndpLLASSLYINIALAGLSILLFVFMtrgldprakLIAVSTILVPVVSIA
                              mdavatAYLGGAVALIVGVAFVWLLYrsldgsphqsalapLAIIPVFAGL

AYLSMFFGIGLtevqvgsem ld                  iyyaryaDWLFTTPLLLLDLALLAkvdrvsIGTLVGVDALMI
AYLAMFFGIGVtevelasgtvld                  iyyaryaDWLFTTPLLLLDLALLAkvdrvtIGTLIGVDALMI
MYLSMLLGYGLtmvpfggeqnp                   iywaryaDWLFTTPLLLLDLALLVdadqgtILALVGADGIMI
SYLGLLSGLtvgmiempaghalag              emvrsqwgrYLTWALSTPMILLALGLLAdvdlgsLFTVIAADIGMC
SYTGLASGLtisvlempaghfaegssvmlggeevdgvvtmwgrYLTWALSTPMILLALGLLAgsnatkLFTAITFDIAMC
SYVGMAYDIGTVIVngnq                       ivglrYIDWLVTTPILVGYVGYAAgasrrsIIGVMVADALMI

VTGLVGALshtp larytWWLFSTICMIVVLYFLATSLraaakergpevastfntLTALVLVLWTAYPILWIIgtegagv
VTGLIGALsktp larytWWLFSTIAFLFVLYYLLtslrsaaakrseevrstfntLTALVAVLWTAYPILWIVgtegagv
GTGLVGALtkvy syrFVWWAISTAAMLYILYVLFFGFTskaesmrpevastfkvlrnVTVVLWSAYPVVWLIGSegagi
VTGLAAAMttsallfrWAFYAISCAFFVVVLSALVtdwaasassagt  aeifdtlrVLTVVLWLGYPIVWAVGVEGLAL
VTGLAAALttsshlmrWFWYAISCACFLVVLYILLVewaqdakaagt  admfntlkLLTVVMWLGYPIVWALGVEGIAV
AVGAGAVVtdgt  lkWALFGVSSIFHLSLFAYLYVIFprvvpdv peqiglfnllknhIGLLWLAYPLVWLFGPAGIGe

vg lgietLLFMVLDVTAKVGFGFILLrsrailgdteapepsagaeasaad
vg lgietLAFMVLDVTAKVGFGFVLLrsrailgeteapepsagadasaad
vp lnietLLFMVLDVSAKVGFGLILLrsraifgeaeapepsagdgaaatsd
Vqsvgvtsway SVLDVFAKYVFAFILLrwvannertvavagqtlgtmssdd
Lp vgvtsWGYSFLDIVAKYIFAFLLLNYLtsnesvvsgsildvpsasgtpadd
ataag VALTYVFLDVLAKVPYVYFFYarrrvfmhsesppapeqatveataad
```

Fig. 1. Aligned sequences of bacterial rhodopsins. Six sequences were aligned using the program **MULTAL**, biased toward the alignment of transmembrane segments (upper-case letters). These were predicted using the method of Jones et al.[20] The sequences are identified by their SWISS-PROT databank identifiers and a rough indication of their relationship is given by the tree to the left.

TABLE I. Membrane Hydrophibicity Data Derived From the Frequency of Occurrence of Residues in the Middle Segment of Single Helix Transmembrane Spanning Segments

A 1.73	C 0.84	D 0.03	E 0.01
F 1.48	G 1.27	H 0.06	I 3.46
K 0.03	L 2.56	M 0.86	N 0.01
P 0.18	Q 0.03	R 0.00	S 0.49
T 0.59	V 2.46	W 0.74	Y 0.59

vation measure to simulate this. The overall similarity (s) between two sequences (j and k) was measured as follows:

$$s_{jk} = \frac{1}{L}\sum_{i=1}^{L} \mathbf{M}_{R_{ij},R_{ik}} \qquad (1)$$

where **M** is a matrix of amino acids relatedness values, R_{ij} is amino acid type index for a residue i in sequence j, and L is the length of the alignment (gaps were ignored). The matrix of similarity values was scaled into the range of $1 \rightarrow m + 1$ as follows:[§]

$$s'_{jk} = 1 + m(s_{jk} - s_{\min})/(s_{\max} - s_{\min}),$$
$$\forall_j = 1, \ldots N, \forall_k = 1, \ldots N; \qquad (2)$$

giving a new set of values s', where $s_{\max}$ and $s_{\min}$ are the maximum and minimum values of all (N by N) pairwise sequence similarities. The scaled values

[§]The symbol "∀" in Eq. (2) should be read: "for all . . . ," and is equivalent to a loop structure in the program that implements the method. The two specifications (for all j, for all k) correspond to two nested loops, processing the whole matrix of s values.

were then used to derive a weighted conservation measure (c) of pairs of residues found at a common position (i) in a multiple sequence alignment (of N sequences), as follows:

$$c_i = \frac{2}{N^2 - N} \sum_{j=1}^{N-1} \sum_{k=j+1}^{N} s'_{jk} \mathbf{M}_{R_{ij},R_{ik}} \quad (3)$$

where R_{ij} is the amino acid type in sequence j at position i.

The preceding measures of hydrophobicity and conservation were combined to produce a score (v) which gave a high value for unconserved hydrophobic positions:

$$v_i = h_i/c_i \quad (4)$$

h is the average hydrophobicity (at position i) and c is the average amino acid similarity obtained from Eq. (3). The quantity v will be referred to as variable-hydrophobicity or, for short, *variphobicity*.

Prediction of membrane spanning segments

Membrane spanning segments were predicted by a novel automatic method that considers the constraint of alternating polarity of each helix as it crosses the membrane (in–out phasing).[20] This method was applied to each sequence in an alignment and a consensus assignment taken when more than 50% of the sequences gave a prediction of helix. Only the exact end-points of the helices remained in doubt but this source of uncertainty was minimized by down-weighting the contribution of the terminal residues in the following calculations by the fraction of sequences predicted as helical (f_i) and the following linear function of the first and last M residues in the sequence:

$$w_i = \begin{cases} \dfrac{i}{M} & \text{if } i < M \\ \dfrac{N-i+1}{M} & \text{if } i > N - M \end{cases} \quad (5)$$

where w is the applied weight and N the number of residues in the helix.

The variphobicity values, $\{v\}$ [Eq. (4)], were shifted by their mean value ($<v>$) to give a new set of values $\{v'\}$ with a mean of zero. The new values were then adjusted in the light of their combined weight ($f_i \cdot w_i$) producing a new mean which was again used to shift the values to maintain a zero weighted-mean. These operations were applied iteratively five times by the following recurrence equation:

$$v_j = v'_j - \frac{1}{N} \sum_{i=1}^{N} f_i w_i v_i, \quad \forall_j = 1, \ldots N. \quad (6)$$

After each cycle, the new values $\{v\}$ become the old values $\{v'\}$ for the next cycle.

Hydrophobic moments

Following Eisenberg and co-workers,[12,13] a hydrophobic moment was calculated for each helix. However, rather than use separate hydrophobic and conservation moments,[23] we calculated a single moment from the weighted variphobic score v [Eq. (6)], as follows:

$$\phi = \tan^{-1} \left[\frac{\Sigma_{i=1}^{N} f_i w_i v_i \cos(ip)}{\Sigma_{i=1}^{N} f_i w_i v_i \sin(ip)} \right] \quad (7)$$

where N is the number of residues in the helix and p the helical periodicity (1.75 radians/residue).

Estimation of helix exposure

A key quantity in determining the packing of helices is their relative degree of exposure to lipid. This is difficult to evaluate in absolute terms as the number and degree of similarity of the sequences in the alignment will set the scale and range of the values of v [Eq. (4)]. However, as a first approximation, it is not unreasonable to assume that roughly half of all the residues (over all helices) will be exposed to lipid. From this, exposed and buried residues can be defined as those with values of v above and below zero, respectively.

Assuming that the buried (and exposed) face of each helix forms a continuous arc when viewed down the helix axis, the size of the arc (2τ) can be calculated simply from the fraction of exposed (variphobic) residues in the helix as

$$\tau = \pi \frac{n}{N} \quad (8)$$

where n is the number of exposed residues and N the total number of residues in the helix.

This estimate of exposure can then be refined using the continuous range of values of v rather than the preceding simple binary classification. Assuming that the estimated variphobic face (of angular size 2τ) is bisected by the variphobic moment, the refined angular size of the face (2Θ) was obtained by varying the value of θ finding the maximum value of following function (see Fig. 2):

$$\max_{\theta = \tau - t \to \tau + t} \left\{ \frac{c + s}{\bar{c} + \bar{s}} \right\}. \quad (9)$$

The quantity s is the sum of the v scores for variphobic residues (those with a $v > 0$) correctly located on the surface face while the quantity $\bar{s}$ is the corresponding sum over incorrectly located residues. The equivalent sums c and $\bar{c}$ correspond to the remaining residues (those with $v < 0$) correctly found in the core sector and incorrectly found outside it, respectively. For example; c was calculated as follows:

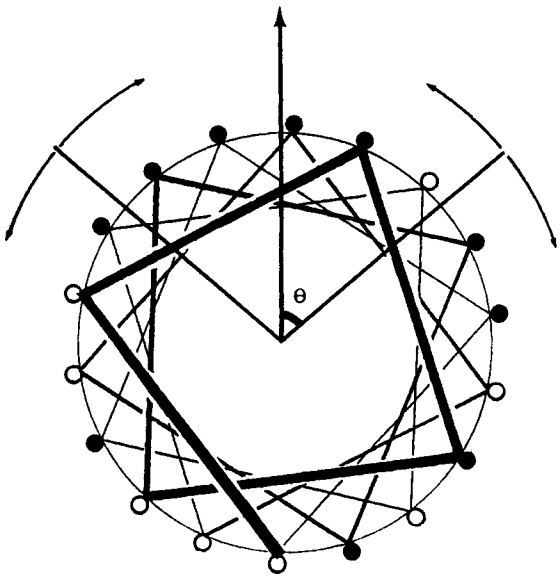


Fig. 2. Refinement of the variphobic arc. Viewed down a helix axis the variable-hydrophobic (variphobic) positions (filled dots), which form a face of the helix, appear as an arc. The arc is bisected by the variphobic moment (radial arrow) and the angle θ defining this arc was varied to optimize the variphobic positions in the arc and the other positions outside. The optimal value of θ is referred to as Θ.

$$c = 1 + \sum_{i=1}^{N} g(v_i). \tag{10}$$

If ω is the angle subtended at the helix axis between residue i and the variphobic moment vector, then the function $g(x)$ was evaluated as

$$g(x) = \begin{cases} |x| & \text{if } \omega > \theta \text{ and } x < 0 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where x is v_i. The corresponding conditions for the remaining three quantities can be found by permuting the two inequalities as follows— $\bar{c}$; $\omega < \theta, x < 0$: s; $\omega < \theta, x > 0$: $\bar{s}$; $\omega > \theta, x > 0$.

The combination of these quantities [Eq. (9)] thus finds the biggest sum of correctly assigned residues $(s + c)$ divided by the sum of incorrect assignments $(\bar{c} + \bar{s})$. The function was maximized over the range of $\tau \pm t$ to avoid fluctuations that can arise with very small or very large values of τ. The value of t was chosen as 60°, allowing the arc to vary in either direction to an extent equivalent to the angular surface occupied by a packed helix.

The inverse of the maximum score provides a convenient measure of the confidence with which an arc has been assigned. It was defined as

$$e = \frac{\bar{c} + \bar{s}}{c + s}, \quad \text{when } \theta = \Theta \tag{12}$$

(where Θ is the value of θ for which e is a minimum). This error term (e) was used to weight the evaluation of how well a helical sequence fits its allocated position in the structure (see "Initial fold evaluation").

Exposure normalisation

For each helix (i), the refined angles Θ_i of predicted variphobic exposure calculated in the previous section might still have an absolute range that depends on the degree of similarity among the original sequences. This, however, can be adjusted by the expected packing density of the helices. In a collection of N packed helices in which none is totally buried, the total number of pairwise packing interactions can be estimated as $4N - 6$.

Allowing 60° ($\pi/3$) for each interaction, the number of packing neighbors can be expressed as an angle and compared with the sum of the refined exposure arcs (Θ) over all helices. This leads to a scaling factor (r) with which the adjust the exposure arcs toward their expected range, as

$$r = \frac{\pi}{3}(2N - 3)/\sum_{i=1}^{N}(\pi - \Theta_i). \tag{13}$$

The additional factor of two (giving $2N - 3$) arises because Θ is a measure of only half the arc (being the displacement from the variphobic moment).

In addition to scaling the estimated arcs, limits can be set on their range. Typically it can be assumed that no helix forms more than five or less than two interactions. These limits were imposed after scaling the arc angles by moving any values outside the range halfway toward their correct limit. The values were then rescaled and recapped five times, so iteratively approaching a set of values in the correct range.

Combinatorics and Construction

Lattice construction

The framework or lattice, over which the chain will be traced, poses few difficulties for membrane proteins with their limited architecture (in contrast to the bewildering variety of forms that must be accommodated for globular proteins). The immediate choice is based on the simple model of a bundle of close-packed cylinders, giving rise to two layers of hexagonally spaced points with a grid spacing of 11 Å between points within the planes and 30 Å between the planes (Fig. 3a).

Theoretically, the two planes might be infinite in extent but as very extended conformations are not expected and computational difficulties increase rapidly with growing size, the lattice was reduced to the minimum required to encompass the anticipated compact forms. For rhodopsin (seven helices), three ranks of 3, 4, and 3 points were considered sufficient.

Membrane helices do not pack exactly longitudinally but are generally inclined slightly relative to each other. This twist is predominantly left-handed

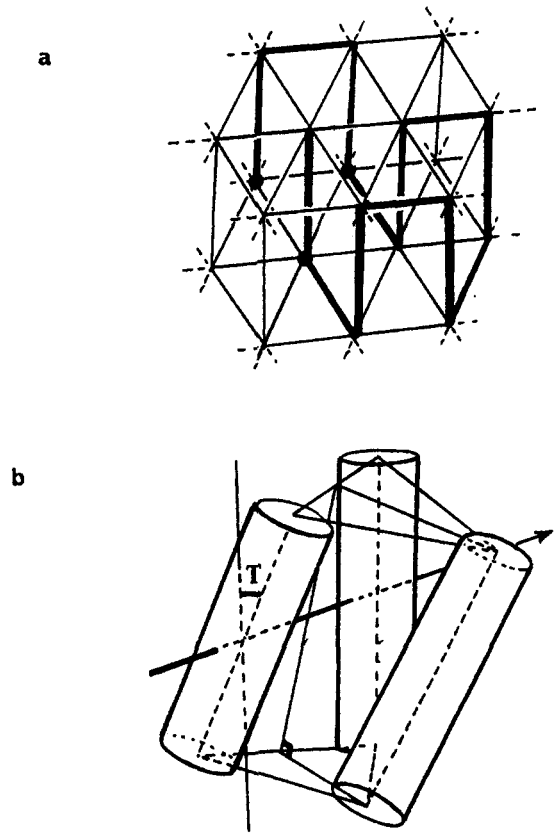


Fig. 3. Lattice construction. **(a)** A simple hexagonal lattice was taken to represent the termini of helices on each face of the membrane. The helix end-points were typically 10–11 Å apart with 30 Å between faces. The fold indicated corresponds to the known structure of bacteriorhodopsin. The three nodes marked by a dot are unique starting points for the combinatorial generation of folds. These generally allow mirror images of most of the folds and occasional repetitions of the same fold. **(b)** Geometric constructs used in the calculation of a twisted lattice. The central axis of the middle rank of the lattice is indicated by an arrow about which a twist angle T is applied between each successive helix. Helices in the next flanking layer of the lattice are then placed such that the separation between the ends of adjacent packing helices remains constant.

and can be modeled by twisting the three layer lattice about the line of the mid points of the central helix axes (Fig. 3b). In practice this can be achieved by taking the first helix axis in the middle layer and twisting the next axis (in the middle layer) to produce a torsion angle (T) when viewed down the line joining the midpoints of both axes (the central axis).

Neighboring points in the flanking layers can be placed such that they are equidistant from their adjacent mid-layer points (as in the untwisted lattice). However, a condition on their placement was that the line connecting them to the mid-point of their neighbors in the middle layer should be orthogonal to the line connecting these midpoints across the membrane. This construction method can be applied recursively to the new set of points so generating a lattice of any required extent.

The relationship between the points in the outer layers is equivalent to the end-points of β-strands in two packed twisted sheets or to the end points of α-helices packed either side of a twisted β-sheet.

Fold generation

Given a set of points, those that lie within a prescribed distance of each other can be taken as possible connections (neighbors). Beginning at any point, each neighbor can be visited in turn with the connection thus made taken to correspond to the path of the polypeptide chain. Applying this procedure recursively, with the condition that each point can be visited only once, leads to the exhaustive enumeration of all windings of the chain over the points (when applied to each unique starting point). This combinatorial method has been previously used on globular proteins,[6,7] and most recently on lattice architectures of the all-α protein class.[25]

The power of this approach depends on how many of the possible folds can be rejected (or avoided) while still retaining the fold corresponding to the native structure. The known structures of integral membrane proteins suggest a range of criteria that can be applied as constraints ranging from almost certain to tentative. Two strong conditions are that the chain can only traverse the membrane in the form of a hydrophobic helix (which in the lattice corresponds to connections between equivalent points on the two faces) and that connections within one face are very unlikely to cross each other. The first condition was imposed absolutely, simply by removing any forbidden transitions from the table of interpoint connections. The second condition, however, was less simple to implement.

Loop crossing constraint

Because of its power in reducing the number of possible folds it was considered importment to implement the loop-crossing check as the folds were being generated, rather than assessing each final fold. This required an algorithm rapidly to check if two line segments (connecting two pairs of points on the same face) intersect each other. Intersection might have been precalculated for all pairs of pairs of points and stored in a table to be accessed during the combinatorial phase of the calculation. This approach requires storage of fourth polynomial order in the number of points [$O(N^4)$], and although feasible, a method which required less storage [$O(N^3)$] was used. This method stored only the sign (S_{ijk}) of the angle of each point (i) relative to the line connecting each pair of points (j, k). For four points (a, b, c, d) defining two line segments a–b and c–d, if $S_{cab} = S_{dab}$ or $S_{acd} = S_{bcd}$, then the line segments do not intersect. To allow for twisted lattices, each S_{ijk} was calculated from the the interpoint vectors, $\mathbf{v}_{ij}$ between points i and j on the same face and $\mathbf{d}_i$ between equivalent points on different faces as

$$S_{ijk} = \text{sgn}\{(\mathbf{v}_{ji} \otimes \mathbf{v}_{jk}) \odot (\mathbf{d}_j + \mathbf{d}_k)\} \quad (14)$$

where $\otimes$ indicates the vector (cross) product, $\odot$ the scalar (dot) product, and the function sgn evaluates the sign of the enclosed expression as $+1$, -1, or 0. The latter value indicates colinear points and when all four points were colinear, intersection was decided by their linear order.

A further condition implemented at the combinatoric stage was to test whether the length of an individual loop between two specified helices was sufficiently long to make the required jump. All other tests were made on the completion of each fold which occurred when all helices had been allocated to a lattice location.

Initial fold evaluation

Sequentially local constraints can be applied as the combinatoric search progresses, however, sequentially long range or global packing aspects are best assessed on completion of a fold. These fall into two groups: interactions that are specific to a particular protein and those that are general. The former were based on specific distance constraints derived from data (or arguments) that do not derive totally from the sequences, while the latter were based on the degree to which the variphobic faces (calculated in "Estimation of helix exposure") can be satisfied in the model.

An assumption in the calculation of variphobic faces was that packing occurs only from one side giving rise to a single exposed face on each helix. This was assessed in each final model by counting the number of neighbours (N) of each helix and checking that between them they had at least $N - 1$ pairwise interactions. Any fold containing a helix with neighbors that failed this test was rejected.

The numbers of neighbors each helix has dictates the expected size of its variphobic arc—from a minimum of 0° (for six neighbors) to a maximum of 300° (for one neighbor). For a helix, i, this expectation (P_i) can be matched to the size of arc calculated from the sequence (Θ_i) and the squared difference, accumulated over all (N) helices, provides a score of the degree to which the sequence fits the fold. This simple measure was modified slightly by multiplication of the error weight (e_i) associated with each arc assignment [Eq. (12)], giving the final measure of fit (F) as follows:

$$F = 10 - \sum_{i=1}^{N} (1 - e_i)\left(P_i - \frac{2\pi\Theta_i}{3}\right)^2. \quad (15)$$

The combinatorially generated folds were then ranked on F.

Model construction

The elaboration of the "stick" models, resulting from the folds traced over the lattice, was carried out as described previously[26] with the only difference being in the orientation of the hydrophobic moments. The moment of each helix was not directed toward the center of gravity (as previously) but *away from* a local center of gravity which was based on the centroid of the lattice end-points of any helices packing against the helix being oriented. This local definition is especially important in the larger problems where the assumption of approximate globularity cannot be made.

TABLE II. Errors in Predicted Bacteriorhodopsin Helix Ends Relative to the Known*

Helix	N	C	Helix	N	C
I	1	3	I	1	3
II	4	4	II	4	4
III	6	0	III	4	0
IV	0	0	IV	0	0
V	1	3	V	1	1
VI	9	0	VI	9	4
VII	3	3	VII	3	3
	+50†			50%+‡	

*The difference (in residues) between the consensus prediction for the amino (N) and carboxy (C) termini of the bacteriorhodopsin helices are tabulated for two consensus definitions.
†Over 50% of the sequences given a transmembrane helical prediction (+50%).
‡50% or more of the sequences give a transmembrane helical prediction (50%+).

RESULTS AND DISCUSSION

Bacteriorhodopsin

Bacteriorhodopsin was taken as the obvious protein on which to test the methods, being relatively small (seven helices) with no large co-factors (such as hemes or chlorophyls) and, although trimeric, has no extensive multimeric or quaternary interactions. Most importantly, it has a known structure solved to reasonable resolution by electron diffraction (3.5 Å in the plane of the membrane).

Helix prediction

The assignment of the seven helices in the sequence alignment was relatively unambiguous in most helices with uncertainty only in the allocation of a few terminal residues (Fig. 1). However, the N-termini of helices III and VI were misassigned by more than one and two turns, respectively (Table II).

The definition of a predicted helix as a consensus of assignments made in "more that 50%" ('+50%', in Table II) of the sequences was relatively insensitive to the precise value of the cutoff in all helices except helix VI where a displacement of a turn (4 residues) would occur at the C-terminus were the cutoff to be trivially redefined as "50% or more" ("50%+," Table

TABLE III. Bacteriorhodopsin Helix Packing Based on the Known Definitions

Helix	Native	Lattice	Helix	I5	M5	I7	M7
I	2.0	2	I	2.2 (.06)	2.1 (.21)	2.2 (.07)	2.6 (.22)
II	3.0	3	II	2.0 (.16)	2.0 (.35)	2.0 (.12)	2.1 (.31)
III	4.5	5	III	4.8 (.09)	4.8 (.41)	4.9 (.10)	4.9 (.42)
IV	2.5	2	IV	3.4 (.18)	2.1 (.25)	3.3 (.17)	2.3 (.25)
V	2.5	3	V	1.9 (.12)	4.2 (.40)	2.0 (.15)	1.9 (.39)
VI	3.5	3	VI	3.6 (.24)	4.1 (.26)	3.7 (.26)	4.9 (.27)
VII	4.0	4	VII	4.1 (.20)	2.6 (.31)	4.0 (.19)	3.3 (.29)
Sum	22.0	22	Difference	4.4	5.7	4.3	6.6
	Observed*			Predicted packing†			

*Observed packings estimated directly from the distance-plot of the known three-dimensional structure (*native*) and from the allocation of the helix end-points onto a regular hexagonal lattice (*lattice*). The packings differ in having only a partial interaction between helices III and V, allowing a corresponding additional partial interaction between helices IV and VI, while in the lattice a full contact is made between III and V with no contact between IV and VI.
†Predicted packings estimated using the known location of the helices in the sequence (but not in space). Two similarity matrices were used: that of Jones et al.[21] (M) and the identity matrix (I). The extent over which the terminal residues were down-weighted was varied between 4 and 6 residues [corresponding to values of M in Eq. (5), of 5 and 7]. Combinations of these are indicated by the column headings I5, I7, M5, and M7. Each value in the table is the number of helices that each helix (I–VII) would ideally like to pack against. This is indicated with an associated fractional error in parentheses [e, in Eq. (12)]. Below each column, a sum-of-squares *difference* between the predicted and the ideal lattice packing is given.

II). To investigate the effect of this sensitivity, both definitions were used below.

Number of packing partners from known helix ends

Calculation of the variphobic arcs produced an estimate of the expected number of packing partners for each helix, along with a measure of confidence in the assignment. These values were compared with those estimated from the molecular model, both in its native form and when imposed onto a regular hexagonal lattice (Table III, see also Fig. 3a).

The predicted packing values for the helices differ slightly depending on the parameters of the method. These were

1. the end-points of the helices (discussed above);
2. the extent of the linear ramp function used to down-weight the contribution of the termini [M in Eq. (5), "Prediction of membrane spanning segments"]
3. the value attached to amino acid differences in closely similar sequences [m in Eq. (2), "Hydrophobicity and conservation"]; and
4. the choice of relatedness (similarity) matrix **(M)** used to measure conservation [Eq. (3), "Hydrophobicity and conservation"].

Fortunately, reasonable choices for these parameters were limited. The sensitivity to the consensus helix definition is only problematic for small numbers of sequences (especially two) but to avoid this complication in the initial characterization of the method, the helix end-points defined by the known structure were used. The ramp function should down-weight at least the first turn of the helix as these would not be subject to the same constraints as those that are more deeply buried in the membrane. Values of $M = 5$ and $M = 7$ (corresponding to 4 and 6 weighted residues) were tried. The immediate choice for the relatedness matrix was that of Jones et al.[18,21] (which had been specifically derived from the class of proteins considered here) but as an alternative, the identity matrix was also considered.

Varying these parameters gave the predicted packings shown in Table III. All schemes correctly identified helix III as the most buried and helices I and II as exposed (less than 3 packing partners). Together helices IV and V were generally more exposed than helices VI and VII (as in the native structure) but this trend was less conserved than the ranking of the three preceding helices.

Visual examination of the predicted arcs of variphobicity on a helical wheel representation suggested that the preceding method might have undervalued the importance of the conservation component in the variphobic score [v in Eq. (94), "Hydrophobicity and conservation"] relative to hydrophobicity. Two different methods were investigated to emphasize conservation. In the simpler measure, the variphobic score was changed from h/c to h/c^2, where h is the average hydrophobicity of a position in a multiple sequence alignment and c the measure of conservation. Alternatively, following the analysis of Baldwin,[3] greater emphasis was at-

tached to differences between similar sequences. The degree to which differences should be weighted introduced another parameter [m, in Eq. (2), "Hydrophobicity and conservation"] for which two values of 4 and 9 were investigated giving a 5-fold and 10-fold emphasis, respectively, to differences in the most conserved sequences.

The use of the squared conservation term in the calculation of the variphobic score led to an improvement in all but one of the schemes investigated. A similar result was obtained for the value of $m = 4$, however, increasing m to 9 resulted in worse deviations from the ideal packing when the amino acid identity matrix was used (see Table V for a summary of these results).

Fold prediction from observed packing

To test the combinatoric generation of folds, the observed packing of the helices was taken as input data. The constraint was imposed that sequentially adjacent helices must pack and that no helix has either one or six partners. Using the ideal packing values assigned from the location of the bacteriorhodopsin helices on a hexagonal lattice (Table II), the correct solution was returned as the highest scoring possibility along with its mirror image (as there is no chirality in the constraints). That this solution should have the best score was predetermined, however, it was not obvious that it should be the only exact solution to the, relatively unspecific, specified packing requirements in over 500 possibilities.

The robustness of this solution was tested by giving a systematic error of ± 1 to the ideal number of packing partners for each helix (without requiring more than five of less than two partners). The results of these trials indicated that the correct solution was reasonably stable, being bettered just twice by alternative folds in the 11 trials (with helix V $+1$ and VI -1) and both times displaced only to second place.

Adding random displacements to the ideal values (evenly distributed in the range $-1 \rightarrow +1$) allowed the stability of the solution to be quantified further. The fraction of correct solutions was calculated for each class range of the sum-of-squares deviation of the displaced packing values from the ideal (Table IV). These results revealed that, above an error of one, the stability of the fold decreased linearly to 50% in the region of 4.0 (helices^2).

Fold selection from predicted packing

The limited exploration of the parameter-space using the known helix end-points as data ("Number of packing partners from known helix ends"), gave little indication any of clearly optimal parameter values. The same range of parameters was therefore retained to investigate the behavior of the model using predicted helix end-points. The known ends, combined with the two alternate predictions gave three sets of data which generated the results given in Table V.

TABLE IV. Fraction of Correctly Predicted Folds Classed by Packing Error*

Class range	Number	% correct
0.0+ → 0.5	2	100.000
0.5+ → 1.0	16	100.000
1.0+ → 1.5	70	91.429
1.5+ → 2.0	101	80.198
2.0+ → 2.5	138	68.116
2.5+ → 3.0	106	57.547
3.0+ → 3.5	75	65.333
3.5+ → 4.0	26	38.462
4.0+ → 4.5	10	50.000
4.5+ → 5.0	6	33.333

The correct fold was taken to correspond to the structure of bacteriorhodopsin imposed on a hexagonal grid, giving the ideal packing partners specified in Table II. Random deviations from these values were generated and measured as a sum-of-squares over the seven helices. In 550 trials, the fraction of correct predictions (% correct*) found in each *class range* was calculated. The *number* of trials in each class is also tabulated.

A reasonable criterion for choosing a parameter set is the stability of the correct prediction under variation of the helix end-point definitions. The closest approach to this ideal was obtained when the scoring scheme was h/c^2 with the matrix of Jones et al. and the four residues at each terminus were down-weighted (a shift of M from 5 to 7 caused only slight deterioration).

In general the predictions using the known helix end-points were poor relative to the predicted ends and of the latter, the consensus of "more than 50%" was correct (or in second place) in most trials.

Allowing Nonadjacent Helices

The loop lengths between the predicted helices were not sufficiently short to constrain the packing of all sequentially neighboring helix pairs in a hairpin conformation. This constraint (which was applied in the results presented above) was, therefore, relaxed such that a loop could connect any two points within 25 Å providing that no other loop was crossed on the same face during the transition. The three helix end-point data sets were used with the best parameter set (Table 5, row 5) to generate new set of folds (Table VI). For comparison, the folds were also generated taking the ideal packings directly as constraints.

Relaxing the loop constraints allowed the number of possible folds to grow to 14,969 and within these, only three can accommodate the ideal helix packings perfectly—the correct fold, its mirror image, and one further fold with many nonlocal connections. The results using the predicted packings do not fall far behind this ideal with their scores all

TABLE V. Predictions Using Different Helix Definitions*

Parameters				Predictions		
Score	m	M	M	Known	50%+	+50%
h/c	0	M	5	5.76 (25)	1.89 (1)	1.40 (1)
h/c	0	M	7	6.58 (24)	1.56 (1)	1.87 (3)
h/c	0	I	5	4.45 (12)	4.00 (15)	3.84 (12)
h/c	0	I	7	4.30 (12)	3.93 (22)	3.65 (9)
h/c^2	0	M	5	3.37 (3)	0.98 (1)	1.20 (1)
h/c^2	0	M	7	3.98 (3)	1.00 (1)	1.92 (3)
h/c^2	0	I	5	3.81 (12)	3.28 (10)	1.78 (1)
h/c^2	0	I	7	5.31 (2)	2.84 (8)	1.78 (1)
h/c	4	M	5	5.77 (25)	7.01 (5)	1.50 (1)
h/c	4	M	7	6.58 (—)	5.83 (8)	0.80 (1)
h/c	4	I	5	3.83 (12)	2.95 (12)	3.41 (12)
h/c	4	I	7	3.86 (6)	3.37 (9)	3.52 (7)
h/c	9	M	5	3.72 (12)	5.15 (9)	1.25 (1)
h/c	9	M	7	3.36 (12)	2.65 (6)	1.60 (1)
h/c	9	I	5	8.30 (—)	5.86 (26)	1.96 (1)
h/c	9	I	7	6.51 (—)	2.06 (1)	2.09 (1)

*Three helix definitions were considered, known, from the known ends, '50%+', from the consensus prediction with "50% or more" assignments taken as helix and "+50%," with a helix prediction in "more than 50%" of sequences (see "Helix prediction"). Each entry includes the sum-of-squares deviation of the predicted packings from the ideal and (in parentheses) the position of the ideal in the ranked list of folds (a "—" indicates a rank of more than 25). This list contains mirror images that cannot be distinguished with the packing score used. The rank of the ideal fold was taken as the better of either alternate forms. The parameters (discussed in "Number of packing partners from known helix ends") were *score*, the form of the variphobic score; m, the weight applied to exaggerate the difference between similar sequences; M, the form of the similarity matrix (either M for that of Jones et al. or I for the identity matrix; and M, which determines the down-weighting of terminal residues (one less than M residues are affected).

TABLE VI. Fold Rankings With Connection Lengths Relaxed*

Helix ends	Fold rank	Group rank	Group size	Group score	Top score
Known	26	5	9	8.1	9.0
Pred 50+	4	2	12	9.2	9.4
Pred +50	14	3	3	9.0	9.7
Ideal	1	1	3	10.0	10.0

*The large number of possible folds (14,969) meant that there were many folds with the same score (referred to as a group). The *fold rank* is the best rank the correct fold can obtain within its group while the *group rank* is the number of groups better than (and including) the group containing the correct fold. *Group size* is the number of equiscoring folds in the group while *group score* is the score of the folds in the group. *Top score* is the score of the best group. The bottom line uses the *ideal* helix packings corresponding to the known structure.

being found in the top five scoring groups giving them a best rank ranging from fourth to twenty-sixth (see Table VI for details).

Model construction

A molecular model was constructed of α-carbon positions using methods developed previously.[25,26] The helices were predicted using the consensus definition "over 50%," while the variphobic arcs were calculated using the designated "best" set of parameters (Table V, row 5). The helix spacing in the construction was varied from 10 to 11 Å and the tilt angle (T) was varied from 0.2 to 0.3 radians. Neither of these variants had any significant effect on the root-mean-square deviation (rmsd) which was measured over all atoms predicted to be helical. With a helix spacing of 10.5 Å and a tilt of 0.25 rad the rmsd was 6.0 Å from the known structure of bacteriorhodopsin (Fig. 4).

Viewed normal to the membrane, the locations of the predicted helices closely approximate their counterparts in the native structure (Fig. 4a). The largest deviation is seen in helix III which on the lattice was allocated the most buried position. Such close packing would leave no space for retinal to bind and the corresponding location in the native structure is displaced toward the surface by roughly 4 Å. In all the helices, the angular orientation corresponds closely to the native, reflecting the good assignment of the variphobic arcs.

As in the native structure itself, the positioning of the predicted helices in the orthogonal direction (Fig. 4b) is much less accurate and is the source of most of the high rmsd. The errors in this dimension arose directly from misprediction of the helix endpoints and range from zero in helix V to over 8 Å in helix III (the amino terminus of which was mispredicted by six residues). This type of error might be corrected by conventional modeling methods[8,11] or by optimal sequence threading.[19]

Rhodopsin family

Helix prediction

The consensus predicted helices for the 30 rhodopsin/opsin sequences agreed remarkably well with the end points identified by Baldwin.[3] Indeed, the majority of differences arose because Baldwin designated a helix of fixed length while the algorithm of Jones[20] allowed limited flexibility in length (Table VII).

Estimated helix packing

Equating the rank order of burial estimated by Baldwin[3] that III > II/VII > VI > I/IV/V, with the expected range in the number of packing partners of 2–5, leads to the following set of estimated packing partners: 2, 4, 5, 2, 2, 3, 4 (for the seven helices, respectively). These values have the same total

Fig. 4. α-Carbon model of the predicted bacteriorhodopsin superposed on the native structure. **(a)**. View down the axis of helix III (the most buried) with the prediction in bold. **(b)** The orthogonal view. In this clearer view some equivalent residues in the central region have been joined with a dashed line.

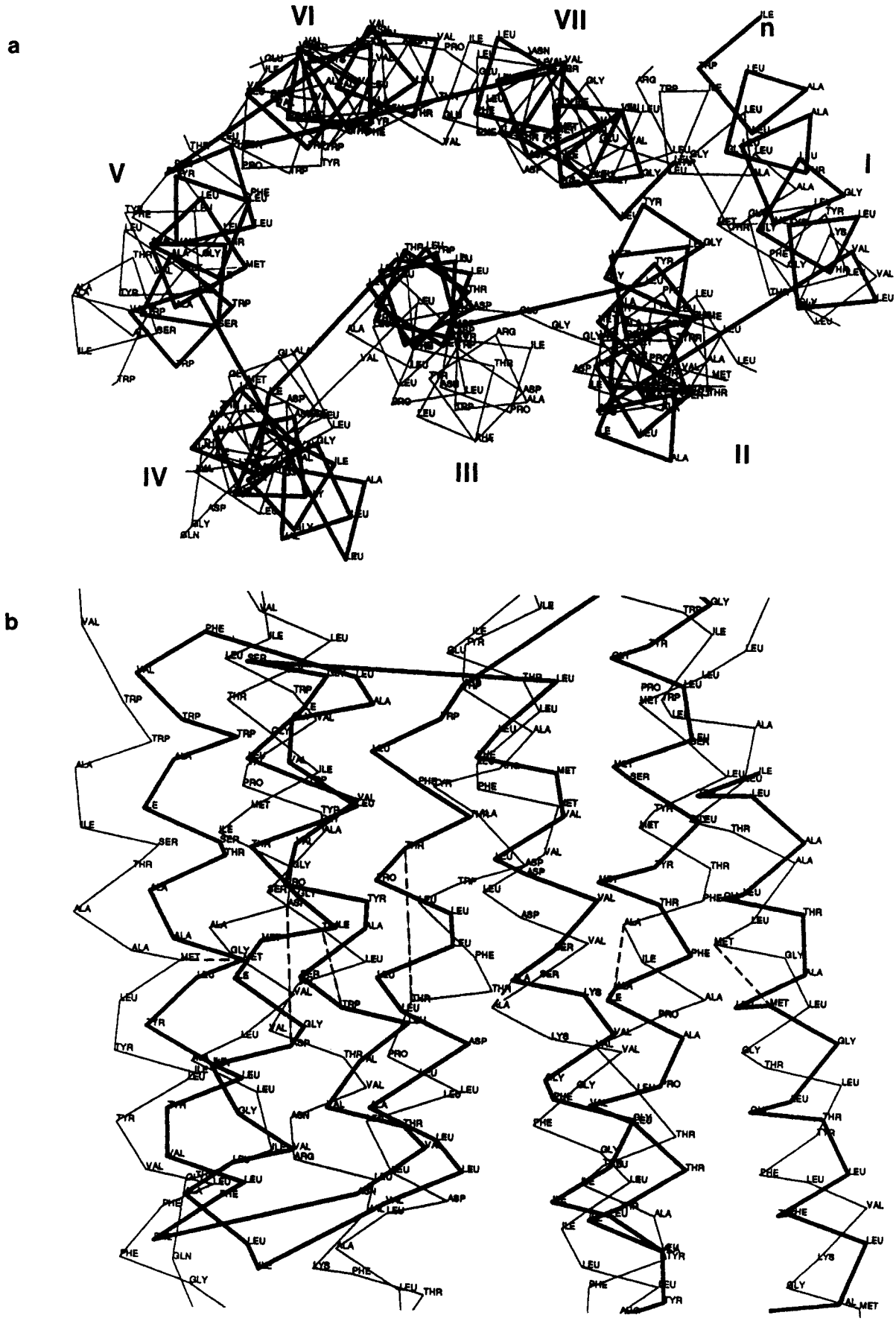


Fig. 4.

TABLE VII. Helix Ends Predicted by Two Methods*

Helix	Jones	Baldwin	Difference
I n	39	38	−1
c	62	63	+1
II n	74	70	−4
c	96	95	+1
III n	114	111	−3
c	133	136	+3
IV n	153	151	−2
c	175	176	+1
V n	203	202	−1
c	223	227	+4
VI n	253	250	−3
c	277	275	−2
VII n	286	286	−0
c	308	311	+3

*The method of Jones et al.[20] was applied to the aligned sequences of the rhodopsin/opsin family. It is compared with the semiautomatic methods of Baldwin[3] (*Baldwin*) based on the extended superfamily of G-protein coupled receptors. The residue numberings refer (arbitrarily) to the sequence of human rhodopsin (SWISS-PROT databank code: OPSD:HUMAN). The final column tabulates the differences in the end-points.

TABLE VIII. Packings Estimated Using the Predicted Rhodopsin Helices*

	h/c^2, $m = 0$†			
Helix	I5	M5	I7	M7
I	2.5 (.03)	2.0 (.16)	2.0 (.14)	2.0 (.13)
II	2.0 (.21)	2.0 (.30)	2.4 (.40)	1.9 (.26)
III	4.7 (.01)	3.1 (.43)	4.3 (.16)	4.0 (.39)
IV	3.9 (.17)	4.9 (.23)	3.7 (.09)	3.9 (.23)
V	1.9 (.10)	2.0 (.23)	2.0 (.09)	2.0 (.21)
VI	2.9 (.22)	2.9 (.26)	3.8 (.13)	2.9 (.26)
VII	4.1 (.14)	5.2 (.13)	3.8 (.10)	5.2 (.13)
Difference	6.01	15.7	5.35	8.03
	h/c, $m = 4$‡			
Helix	I5	M5	I7	M7
I	2.0 (.08)	1.9 (.21)	2.0 (.08)	1.9 (.20)
II	3.6 (.38)	2.0 (.52)	2.0 (.39)	1.9 (.36)
III	4.2 (.21)	1.9 (.54)	4.4 (.19)	1.9 (.57)
IV	3.2 (.08)	4.5 (.23)	3.3 (.09)	4.8 (.24)
V	2.0 (.09)	1.9 (.29)	2.0 (.08)	1.9 (.29)
VI	3.4 (.14)	4.8 (.35)	3.5 (.13)	4.6 (.33)
VII	3.6 (.12)	5.0 (.21)	3.8 (.12)	5.0 (.22)
Difference	3.66	22.0	3.37	23.0

*See footnotes to Table III for details.
†Using the variphobic score emphasized by using conservation squared.
‡Using conservation weighting by the contribution from similar sequence pairs with $m = 4$ in Eq. (2).

amount of packing as both the native and lattice estimates for bacteriorhodopsin (Table IIIa) and differ from the estimate based on the native structure by 2.5 and from the ideal (lattice) values by 2. It can be estimated from Table IV that a set of values with this much a deviation from the ideal should predict the fold corresponding to the native with 70–80% confidence. When used as constraints, the ideal fold corresponding to the native structure of bacteriorhodopsin was selected as best.

Predicted helix packing

Little guidance toward any optimal parameter set was obtained from the investigations with bacteriorhodopsin except that some emphasis of conservation over the basic variphobic scoring scheme seemed preferable. Two methods were applied to the rhodopsin/opsin data: (1) using the squared conservation measure (h/c^2) and (2) weighted differences according to sequence similarity [$m = 4$ in Eq. (2)]. Allowing both the similarity matrix and the extent of damping of the terminal residues to vary as before, the results given in Table VIII were obtained.

Using the identity matrix, the results for the rhodopsin family were comparable to the quality of fit to the ideal packings seen with bacteriorhodopsin. Within this group the results generated by the method using increased conservation weighting between similar sequences was best.

When the similarity matrix was used, however, the estimated packing for the helices was wildly wrong. A possible reason for this unexpected result might be that the effect of the emphasis given to changes between more similar sequences is linked with the nature of the similarity matrix: for example—any pair of matched residues in the alignment of two very similar sequences will be equivalent under the identity matrix but using a relatedness matrix, the relative mutabilities of the residues will give rise to different values. Emphasizing this difference by the overall similarity of the sequences will have no effect when the identity matrix is used but will exaggerate the mutability of the residue when using a similarity matrix. The effect is most acute in the assessment of highly conserved positions of relatively mutable residues (such as the aliphatic hydrophobics) which were common in the rhodopsin alignment.

Relaxing hairpin packing

The preceding rhodopsin results had been generated with the constraint that sequentially adjacent helices should pack as hairpins. This constraint was relaxed under the same conditions for bacteriorhodopsin (above) and the helix packings were estimated using the parameters specified in the first column of Table VIII (I5). Under these conditions the correct fold lay in a group of equiscoring folds ranked 107–131 in the list of 14,969 folds.

From the alignment of rhodopsin/opsin sequences considered here, it would reasonable to assume that the loops between helices I–II and VI–VII were short

enough to constrain a hairpin packing conformation of the flanking helices. Applying these two constraints to the combinatoric fold generations resulted in the improved position for the correct fold in a group spanning 70–79 in the ranking list.

Additional constraints

Given the reasonable assessment of helix packing for rhodopsin obtained through the use of the identity matrix and the constraint of hairpin packing between adjacent helices, four unique folds scored better than the fold corresponding to bacteriorhodopsin (which has been assumed to correspond to rhodopsin). Ignoring the finer discriminations between mirror image structures, the biochemical arguments used by Baldwin[3] would be sufficient to eliminate all of the four incorrect structures.

Specifying that helices III and helix VII should pack to form a salt-bridge between lysine 296 and glutamic acid 113 in the human rhodopsin sequence (equivalent to lysine 216 and aspartic acid 85 in bacteriorhodopsin Henderson et al.,[16] their Fig. 18) eliminates one possibility while the others can be discounted on the grounds that helix III is not sufficiently close to form a disulfide bridge to the loop between helices V and VI (residues 110 and 187 in the human rhodopsin sequence). It is also unlikely that models (3) and especially (2) would have sufficient space to bind the chropophore (Fig. 5).

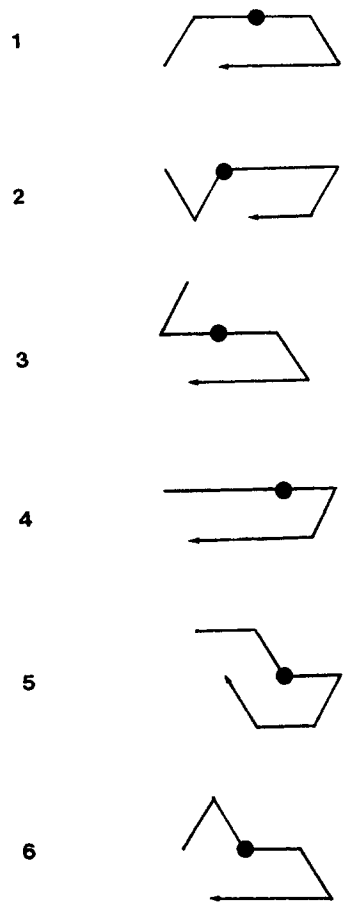


Fig. 5. High scoring alternative folds. Folds are depicted on a simple hexagonal lattice that score highly using the rhodopsin/opsin sequences with the identity matrix to score amino acid conservation. The folds begin at the terminus marked by a dot. Only four folds score better than the fold (5) that is assumed to correspond to rhodopsin. Fold (4) could be excluded by the remote packing of helices III and VII, while the others (1–3) might be unable to form the disulfide bond between helix III and the loop between helices IV and V.

CONCLUSIONS

The prediction of the structure of integral membrane proteins from sequence is of equal importance to the prediction of globular protein structure. Despite being, in principle, a more tractable problem, few methods have been developed that are not based on the use of a "homologous" structure and in many applications this approach has been followed even though there was no significant sequence similarity between the template sequence and that being modeled. An exception is the approach of Baldwin[3] to the G-protein coupled receptors which avoided any assumption of similarity to the bacterial rhodopsin family. This approach, however, is only partially automated and still relies on some biological insight on the part of the investigator to identify functionally equivalent subfamilies. The methods developed here are, by contrast, completely automatic in the entire progression from sequence databank searching to the construction of a three-dimensional model. This automation has resulted in some aspects of Baldwin's method being streamlined—in particular the analysis of subfamily variability. This loss has been considered acceptable given the benefits that are gained by allowing the "blind" application of the method.

The key to the method described here is the estimation of the size of the variable-hydrophobic (variphobic) face of the helix. Using ideal data it was found that the size of these faces must be predicted with high accuracy to give a good chance of the prediction finding the correct fold among the many (combinatorially generated) alternatives. A sum-of-squares deviation from the ideal packing of less than 3–4 (helices^2/face) over the seven helices (0.25 root-mean-square) was found to be required to have better than a 50% chance of getting the correct answer. This accuracy cannot be guaranteed, however, given the unavoidable fluctuation in the predicted endpoints for the transmembrane helices and uncertainty in the choice of parameters used to measure sequence hydrophobicity and conservation. Typically, about six (unique) alternative folds would need to be considered to ensure a reasonable chance of including the correct result.

Our method was applied to two sequence alignments which differed markedly in both extent and

degree of internal similarity among the sequences. There were only six bacterial sequences which (with the exception of a pair of archaerhodopsin sequences) were reasonably diverse. By contrast, the eukaryote rhodopsin/opsin family of 30 sequences (taken directly from the sequence databank with no selection) included many closely related subgroups. An aspect of our method, revealed by its application to these two very different data sets was that the best choice of parameters is probably a function of the number and degree of similarity of the sequences included in the original alignment. The results with the bacterial sequences suggested that a small number of distantly related sequences were best analyzed using a similarity (or relatedness) matrix, with no exaggeration of the differences between similar sequences. By contrast, using the eukaryotic rhodopsin/opsin sequences, the similarity matrix gave very poor results while the identity matrix gave good results. In this initial characterization of the method, this aspect has not been fully investigated and any further work on other members of the G protein coupled receptor family must await the full solution of the rhodopsin structure.

Further application of the method to other protein families is limited only by the condition that the helices are close-packed, and reasonably normal to the plane of the membrane. Under these conditions, the variphobic arcs, on which the success of the method depends, should be present. All other aspects, such as the assumption in the current work of a hexagonal lattice or the number of helices and their permitted connectivities, can be altered.

ACKNOWLEDGMENTS

Tom Flores is thanked for useful discussion and Joyce Baldwin for valuable comments on the manuscript. D.T.J. was supported by a Welcome Trust Biomathematical Fellowship.

REFERENCES

1. Bairoch, A. PC/Gene: a protein and nucleic acid sequence analysis microcomputer package, PROSITE: a dictionary of sites and patterns in proteins, and SWISS-PROT: a protein sequence data bank. Ph.D. thesis, University of Geneva, 1990.
2. Bairoch, A., Boeckmann, B. The SWISS-PROT protein sequence data bank. Nucleic Acids Res. 19:2247–2249, 1991.
3. Baldwin, J. M. The probable arrangement of the helices in G protein-coupled receptors. EMBO J. 12:1693–103, 1993.
4. Barlow-Close, D., Robbins, A. D., Rubin, P.H., Stallman, R. The GAWK manual. Technical report, Free Software Foundation, 1989. Published by the Free Software Foundation, 675 Massachusetts Ave., Cambridge, MA, USA.
5. Blanck, A., Osterhelt, D. The halo-opsin gene: II. Sequence, primary structure of halorhodopsin an comparison with bacteriorhodopsin. EMBO J. 6:265–273, 1987.
6. Cohen, F. E., Sternberg, M. J. E., Taylor, W. R. Analysis and prediction of protein β-sheet structures by a combinatorial approach. Nature (London) 285:378–382, 1980.
7. Cohen, F. E., Sternberg, M. J. E., Taylor, W. R. Analysis and prediction of the packing of α-helices against a β-sheet in the tertiary structure of globular proteins. J. Mol. Biol. 156:821–862, 1982.
8. Cronet, P., Sander, C., and Vriend, G. Modelling of transmembrane 7 helix bundles. Prot. Eng. 6:59–64, 1993.
9. Devereux, J., Haeberli, P., Smithies, O. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. 12:387–395, 1984.
10. Donnelly, D., Johnson, M. S., Blundell, T. L., Saunders, J. An analysis of the periodicity of conserved residues in sequence alignments of g protein-coupled receptors: Implications for the three dimensional structure. FEBS Lett. 251: 109–116, 1989.
11. Donnelly, D., Overington, J. P., Ruffle, S. V., Nugent, J. H. A., Blundell, T. L. Modelling α-helical transmembrane domains: the calculation and use of substitution tables for lipid-facing residues. Protein Sci. 2:55–70, 1993.
12. Eisenberg, D., Weiss, R. M., Terwilliger, T. C. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. Nature (London) 299:371–374, 1982.
13. Eisenberg, D., Weiss, R. M., Terwilliger, T. C. The hydrophobic moment detects periodicity in protein hydrophobicity. Proc. Natl. Acad. Sci. U.S.A. 81:140–144, 1984.
14. Green, N. M. Biological membranes: the semiotics of charge. Nature (London) 351:349–350, 1991.
15. Green, N. M., Taylor, W. R., Brandl, C. J., Korczak, B., MacLennan, D. H. A. Structural and mechanistic implications of the amino acid sequence of calcium transporting ATPases. In "Calcium and the Cell," vol. 122 of Ciba Found. Symp. Evered, C., ed. New York: J. Wiley, 1986: 93–113.
16. Henderson, R., Baldwin, J. M., Ceska, T. A., Zemlin, F., Beckman, E., Downing, K.H. Model for the structure of bacteriorhodopsin based on high-resolution electron cryomicroscopy. J. Mol. Biol. 213:899–929, 1990.
17. Jennings, M. J. Topography of membrane proteins. Annu. Rev. Biochem. 58:999–1027, 1989.
18. Jones, D. T. Structural approaches to protein sequence analysis. Ph.D. thesis, Department of Biochemistry and Molecular Biology, University College, University of London, 1993.
19. Jones, D. T., Taylor, W. R., Thornton, J. M. A new approach to protein fold recognition. Nature (London) 358: 86–89, 1992.
20. Jones, D. T., Taylor, W. R., Thornton, J. M. A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry, in press.
21. Jones, D. T., Taylor, W. R., Thornton, J. M. A mutation data matrix for transmembrane proteins. FEBS Lett., in press.
22. Pardo, L., Ballesteros, J. A., Osman, R., Weinstein, H. On the use of the transmembrane domain of bacteriorhodopsin as a template for modelling the 3-D structure of the G-protein coupled receptors. Proc. Natl. Acad. Sci. U.S.A. 89: 4009–4012, 1992.
23. Rees, D. C., Deantonio, L., Eisenberg, D. The hydrophobic organisation of membrane proteins. Science 245:510–513, 1989.
24. Taylor, W. R. A flexible method to align large numbers of biological sequences. J. Mol. Evol. 28:161–169, 1988.
25. Taylor, W. R. Towards protein tertiary fold prediction using distance and motif constraints. Prot. Eng. 4:853–870, 1991.
26. Taylor, W. R. Protein fold refinement: building models from idealised folds using motif constraints and multiple sequence data. Prot. Eng. 6:593–596, 1993.
27. Traxler, B., Boyd, D., Beckweith, J. The topological analysis of integral cytoplasmic membrane proteins. J. Memb. Biol. 132:1–11, 1993.
28. von Heijne, G. Membrane-protein structure prediction: Hydrophobicity analysis and the positive-inside rule. J. Mol. Biol. 225:487–494, 1992.