J. Mol. Biol. (1984) **179**, 125–142

Analysis of Membrane and Surface Protein Sequences with the Hydrophobic Moment Plot

D. EISENBERG, E. SCHWARZ, M. KOMAROMY AND R. WALL

Molecular Biology Institute, Departments of Chemistry and Biochemistry and Microbiology and Immunology
University of California, Los Angeles
Calif. 90024, U.S.A.

(Received 17 January 1984, and in revised form 11 June 1984)

An algorithm has been developed which identifies α-helices involved in the interactions of membrane proteins with lipid bilayers and which distinguishes them from helices in soluble proteins. The membrane-associated helices are then classified with the aid of the *hydrophobic moment plot*, on which the hydrophobic moment of each helix is plotted as a function of its hydrophobicity. The magnitude of hydrophobic moment measures the amphiphilicity of the helix (and hence its tendency to seek a surface between hydrophobic and hydrophilic phases), and the hydrophobicity measures its affinity for the membrane interior.

Segments of membrane proteins in α-helices tend to fall in one of three regions of a hydrophobic moment plot: (1) monomeric transmembrane anchors (class I HLA transmembrane sequences) lie in the region of highest hydrophobicity and smallest hydrophobic moment; (2) helices presumed to be paired (such as the transmembrane M segments of surface immunoglobulins) and helices which are bundled together in membranes (such as bacteriorhodopsin) fall in the adjacent region with higher hydrophobic moment and smaller hydrophobicity; and (3) helices from surface-seeking proteins (such as melittin) fall in the region with still higher hydrophobic moment. α-Helices from globular proteins mainly fall in a region of lower mean hydrophobicity and hydrophobic moment.

Application of these methods to the sequence of diphtheria toxin suggests four transmembrane helices and a surface-seeking helix in fragment B, the moiety known to have transmembrane function.

1. Introduction

Rapid DNA sequencing methods have provided the amino acid sequences of numerous membrane proteins. This has stimulated several computational methods for detection of transmembrane α-helices within these amino acid sequences (e.g. see Segrest & Feldman, 1974; Engelman *et al.*, 1982; Kyte & Doolittle, 1982; Argos *et al.*, 1982). The more recent of these methods are based on a moving "window", 7 to 27 residues in length, which passes along the amino acid sequence. The mean hydrophobicity of the residues within the window is recorded for each position of the window, and the most hydrophobic segments are

0022-2836/84/290125-18 $03.00/0

presumed to correspond to transmembrane segments. However, the problem of what distinguishes such segments from hydrophobic segments of soluble proteins has not been considered.

Here we expand previous approaches in two ways. (1) We report a variation of the moving window algorithm which detects probable transmembrane helices in the sequences of nearly all known membrane proteins and which also correctly fails to detect transmembrane sequences in known soluble proteins. (2) We classify membrane-related helices, using both their hydrophobicities and their hydrophobic moments displayed in a hydrophobic moment plot. The co-ordinates of a helix on this plot may tend to reflect whether it is a membrane anchor, or is a helix that is paired or bundled within the membrane.

2. Methods

(a) *Hydrophobicities*

In all calculations, the amino acid sequence of a protein is represented as a sequence of residue hydrophobicities, one number assigned to each type of amino acid. The numerical values we use are the consensus scale of Eisenberg *et al.* (1982*b*), normalized so that the mean value of the hydrophobicities is zero and the standard deviation is unity. The values are given in Table 1.

TABLE 1

Hydrophobicity scales used in membrane studies

Residue	Consensus†	Normalized consensus‡
Arginine	−1·76	−2·53
Lysine	−1·10	−1·50
Aspartic acid	−0·72	−0·90
Glutamine	−0·69	−0·85
Asparagine	−0·64	−0·78
Glutamic acid	−0·62	−0·74
Histidine	−0·40	−0·40
Serine	−0·26	−0·18
Threonine	−0·18	−0·05
Proline	−0·07	0·12
Tyrosine	0·02	0·26
Cysteine	0·04	0·29
Glycine	0·16	0·48
Alanine	0·25	0·62
Methionine	0·26	0·64
Tryptophan	0·37	0·81
Leucine	0·53	1·06
Valine	0·54	1·08
Phenylalanine	0·61	1·19
Isoleucine	0·73	1·38

† From Eisenberg *et al.* (1982*b*) based on 5 other scales.

‡ The normalized consensus values have been scaled to have a mean of 0·00 and a standard deviation of 1·00.

(b) *Hydrophobic moments*

The value of the hydrophobic moment of an idealized α-helix of N residues, in which side-chains protrude perpendicular to the helix axis at regular 100° intervals, is given by:

$$\mu_H = \left\{ \left[\sum_{n=1}^{N} H_n \sin(\delta n) \right]^2 + \left[\sum_{n=1}^{N} H_n \cos(\delta n) \right] \right\}^{1/2}, \quad (1)$$

in which H_n is the hydrophobicity of the nth residue and δ is 100° (Eisenberg *et al.*, 1982*a*). The value of μ_H is a measure of the amphiphilicity of the helix. α-Helices such as melittin, having mainly polar side-chains protruding from one side and mainly apolar side-chains protruding from the other, are characterized by large values of μ_H and tend to seek surfaces between polar and apolar phases, such as the surfaces of membranes. Equation (1) can be generalized to describe other periodic structures by permitting other values of δ (Eisenberg *et al.*, 1984).

(c) *Detection of transmembrane helices in amino acid sequences*

As noted by others (Segrest & Feldman, 1974; Henderson, 1979; Engelman & Steitz, 1981; Kyte & Doolittle, 1982; Argos *et al.*, 1982), many membrane-associated proteins contain hydrophobic segments, often 18 to 24 residues in length, that are probably membrane-penetrating α-helices. Such a length is thought to be appropriate for spanning a membrane because 21 residues coiled into an α-helix approximate the thickness of the apolar portion of a lipid bilayer. Also, the fully hydrogen-bonded backbone of the α-helix is more likely to seek the apolar part of the membrane than non-hydrogen-bonded conformation. Similarly, Eisenberg *et al.* (1982*a*) have argued that amino acid segments that are highly amphiphilic when arranged as an α-helix probably seek the surface between membrane and aqueous phases.

Probable transmembrane helices in a protein are identified by the following procedure.

(1) *Locating candidate transmembrane segments.* A 21-residue "window" is run along the entire protein sequence, generating a succession of 21-residue segments. Determining which of these segments are genuine transmembrane sequences consists of disqualifying most or all of them by: (a) eliminating all with mean hydrophobicity $\langle H \rangle$ less than 0·42 (a value between those for cysteine and glycine); (b) selecting as a "candidate" from the remaining segments the one with the highest $\langle H \rangle$ value, and then disqualifying all other segments with one or more residues in common with the candidate; and (c) repeating the process of selection and disqualification in (b) until all remaining non-disqualified segments have been selected as candidates or disqualified.

(2) *Establishing the status of candidates.* All candidates are themselves rejected at this point unless (a) there is at least one candidate with $\langle H \rangle$ greater than or equal to 0·68 or (b) there are two candidates whose summed $\langle H \rangle$ values is greater than or equal to 1·10. If either of these conditions exists, all candidates of the protein are then considered actual transmembrane sequences. Below we call these especially hydrophobic helices that are necessary for membrane penetration "initiators", and note that they imply some co-operativity in the folding of proteins into the membrane.

In selecting candidates for transmembrane helices, additional steps are needed if the hydrophobicity profile has 2 or more maxima. Where 2 or more 21-residue windows on a protein sequence have equally high $\langle H \rangle$ values, the maximum with neighboring windows having the largest mean $\langle H \rangle$ value is selected. Should 2 or more maxima exist whose neighbors have the same mean $\langle H \rangle$ value, the neighbors preceding and succeeding the maxima by 2 residues are then examined and the maximum with the highest mean $\langle H \rangle$ value of such neighbors is selected.

(3) *Tentative characterization of membrane-associated segments.* Transmembrane sequences, once selected, are classified on a hydrophobic moment plot. As explained in Results, sequences from proteins of different functional type fall in different regions of the plot. The steps of this procedure are: (a) running a window of 11 residues through the sequence,

generating a set of successive windows with values of $\langle H \rangle$ and mean hydrophobic moment $\langle \mu_H \rangle$, and (b) selecting the one window with the highest $\langle \mu_H \rangle$. This window's $\langle H \rangle$ and $\langle \mu_H \rangle$ are placed on the hydrophobic moment plot.

In practice, a sequence can yield $\langle \mu_H \rangle$ values having more than one maximum. Where 2 or more maxima exist, preference goes to the one whose neighboring windows have the highest mean $\langle \mu_H \rangle$ value. Windows at the end of a membrane-associated sequence are considered to have only one neighbor, that neighboring window within the sequence. This prevents contamination by data from globular domains of a membrane protein. In rare cases even this fails to discriminate between 2 or more windows with maximum values of $\langle \mu_H \rangle$. In this case the one with the highest $\langle H \rangle$ value is chosen.

(4) *Defining regions of membrane association on the hydrophobic moment plot.* Regions on the hydrophobic moment plot were associated empirically with various types of protein–membrane interactions from the pattern of data points generated by the preceding steps.

The outer boundary of the plot is defined by calculated hydrophobic moments for a set of model peptides composed of isoleucine and arginine. For any given mean hydrophobicity, the boundary shows the maximum possible hydrophobic moment. The internal divisions were set empirically, and are proposed as rough indicators rather than as firm phase boundaries between classes of proteins.

(d) *Amino acid sequences of membrane proteins*

Amino acid sequences of some 36 membrane proteins were investigated. These include all proteins of Table 2A plus the following. Subunits I, II and III of *Homo sapiens* mitochondrion (Anderson *et al.*, 1982); subunits I, II and III of *Saccharomyces cerevisiae* (Dayhoff *et al.*, 1982); and hypothetical proteins 1, 2, 3, 4, 4L, 5, 6 and A6L of *H. sapiens* mitochondrion (Dayhoff *et al.*, 1982).

3. Results

(a) *Helices in transmembrane proteins*

Transmembrane α-helices were detected by the algorithm described in Methods from the known amino acid sequences of 36 membrane-related proteins. The algorithm implies that an "initiator" is necessary for transmembrane association. This initiator is either a very hydrophobic single helix (mean hydrophobicity $\langle H \rangle \geq 0{\cdot}68$), or a moderately hydrophobic pair of helices (whose $\langle H \rangle$ values sum to $\pm 1{\cdot}10$). Other helices within the same polypeptide are then accepted as transmembrane if they are above a threshold hydrophobicity ($\langle H \rangle \geq 0{\cdot}42$). This algorithm detects seven hydrophobic helices in bacteriorhodopsin, where seven are known to exist (Henderson & Unwin, 1975). The choice of threshold hydrophibicity values (0·68 and 0·42) was determined by the optimal separation of soluble and transmembrane proteins. These values are arbitrary in that they have not been determined independently and depend entirely on the hydrophobicity scale used here. The residue composition of the putative transmembrane helices is largely hydrophobic; however, there are some charged residues and a good deal of threonyl and seryl residues.

Such hydrophobic helices are not detected in the amino acid sequences of globular and surface-seeking proteins, as shown in Table 2B and C. Diphtheria toxin, an apparent exception, is discussed below. The globular proteins are controls, selected for their binding of hydrophobic molecules (globins, apolipoproteins, serum albumin), their large size (β-galactosidase, ribulosbis-

TABLE 2

Transmembrane and surface-seeking segments of proteins, identified from amino acid sequences

Protein name and reference	Residues in transmembrane sequence Start	End	% MEM	% SURF (non-MEM)
A. *Membrane proteins*				
Bacteriorhodopsin, *Halobacterium halobium*	9	29	42	3
Khorana *et al.* (1979)	42	62		
	83	103		
	107	127		
	136	156		
	179	199		
	203	223		
Rhodopsin, bovine	39	59	30	3
Dratz & Hargrave (1983)	75	95		
	113	133		
	153	173		
	203	223		
	255	275		
	286	306		
Acetylcholine receptor, *Torpedo californica*				
α-Subunit	−23	−3	20	5
Noda *et al.* (1982)	213	233		
	245	265		
	277	297		
	408	428		
β-Subunit	−18	3	15	5
Noda *et al.* (1983*b*)	219	239		
	251	271		
	284	304		
	438	458		
γ-Subunit	220	240	17	4
Noda *et al.* (1983*a*)	254	274		
	287	307		
	450	470		
δ-Subunit	−20	1	15	3
Noda *et al.* (1983*b*)	226	246		
	254	274		
	292	312		
	456	476		
Lactose carrier protein, *E. coli*	14	34	39	5
Foster *et al.* (1983)	46	66		
	76	96		
	103	123		
	145	165		
	167	187		
	219	239		
	263	283		
	291	311		
	312	332		
	349	369		
	380	400		

TABLE 2 *(continued)*

Protein name and reference	Residues in transmembrane sequence		% MEM	% SURF (non-MEM)
	Start	End		
Chlorophyll-binding protein, *Lemna gibba*	64	84	13	4
E. M. Tobin *et al.* (personal communication)	116	136		
	184	204		
ATP operon, *E. coli*				
Gene product 1	19	39	53	4
Gay & Walker (1981)	42	62		
	90	110		
Gene product 2	40	60	39	9
Gay & Walker (1981)	70	90		
	101	121		
	122	142		
	146	166		
	182	202		
	217	237		
	242	262		
Gene product 3	12	32	57	13
Gay & Walker (1981)	53	73		
Gene product 4	12	32	12	1
Gay & Walker (1981)				
Fumarate reductase, *E. coli*				
M_r 15,000 peptide	30	50	30	2
Grundstron & Jaurin (1982)	63	83		
	109	129		
M_r 13,100 peptide	26	46	49	7
Grundstron & Jaurin (1982)	56	76		
	99	119		
Serine sensory transducer, *E. coli*	10	30	8	7
Boyd *et al.* (1983)	194	214		
	361	381		
Cytochrome P450, rabbit phenobarbital-induced	3	23	6	6
	94	114		
Heinemann & Ozols (1982)	162	182		
	199	219		
	285	305		
	443	463		
Histidine transport system, *S. typhimurium*				
Gene product M	27	47	22	16
Higgins *et al.* (1982)	58	78		
	105	125		
	158	178		
	200	220		
Gene product P	—	—	2	5
Higgins *et al.* (1982)				
Gene product Q	13	33	36	8
Higgins *et al.* (1982)	59	79		
	92	112		
	153	173		
	195	215		

TABLE 2 *(continued)*

Protein name and reference	Residues in transmembrane sequence		% MEM	% SURF (non-MEM)
	Start	End		
Respiratory NADH dehydrogenase, *E. coli* Young *et al.* (1981)		—	4	3
Procoat, bacteriophage M13 Wickner (1980)	4 45	24 65	32	9
Leader peptidase Wolfe *et al.* (1983)	2 64 90 283	22 84 110 303	8	0
B. *Surface-seeking proteins*				
Cecropin A, *Hyalophoria cecropia* Steiner *et al.* (1981)		—	15	22
Cecropin B, *H. cecropia* Steiner *et al.* (1981)	—	—	22	33
δ-Hemolysin, *S. aureus* Fitton *et al.* (1980)		—	0	100
δ-Hemolysin, *S. aureus* (canine strain) Fitton *et al.* (1980)		—	0	94
Artificial cytotoxin DeGrado *et al.* (1981)	—	—	0	69
Artificial cytotoxin W. F. DeGrado (personal communication)	—	—	0	100
Mellitin, *Apis mellifera* Habermann (1972)	—	—	19	23
Mellitin, *Apis florea* Kreil (1973)		—	13	57
C. *Globular proteins*				
Myoglobin, *Physeter catadon* Feldman (1976)	—	—	2	9
Hemoglobin, β-subunit, *Homo sapiens* Feldman (1976)	—	—	4	5
Triose phosphate isomerase Feldman (1976)		—	7	0
Concanavalin A Feldman (1976)	—	—	1	3
Citrate synthase Bloxham *et al.* (1981)		—	3	5
D-Ribulose-1,5-bisphosphate carboxylase/oxygenase, *Nicotiana tabacum* Large subunit Shinozaki & Sugiura (1982)		—	2	4
Small subunit Muller *et al.* (1983)	—	—	0	0
Serum albumin, *Bos taurus* Brown & Shockley (1982)		—	1	4

TABLE 2 *(continued)*

Protein name and reference	Residues in transmembrane sequence: Start	End	% MEM	% SURF (non-MEM)
Apolipoproteins *H. sapiens*				
A-I Brewer *et al.* (1978)			0	0
A-II Brewer *et al.* (1972)			0	10
C-I Shulman *et al.* (1975), Jackson *et al.* (1974)			0	0
C-II Jackson *et al.* (1977)			0	0
C-III Brewer *et al.* (1974)			0	0
Prophospholipase A2, *Sus porcus* Dayhoff *et al.* (1982)			0	0
Phosphorylase, rabbit Dayhoff *et al.* (1982)			1	6
β-Galactosidase, *E. coli* Dayhoff *et al.* (1982)			0	1
Ribitol dehydrogenase, *Enterobacter aerogenes* Dayhoff *et al.* (1982)			10	7
Lactate dehydrogenase, dogfish Dayhoff *et al.* (1982)			5	5
Diphtheria toxin, Coryne bacteriophage Greenfield *et al.* (1983)	−21 269 301 338 418	−1 289 321 358 438	8	4
Crambin Hendrickson & Teeter (1981)			25	0

% MEM is the percentage of all 11-residue windows in the sequence which fall in the "transmembrane" region of Fig. 1. % SURF (non-MEM) is the percentage of those 11-residue windows in the sequence which fall in the "surface" region, excluding from the total the number that fall in the "transmembrane" region.

phosphate carboxylase, phosphorylase, etc.), implying perhaps small surface to volume ratios and highly hydrophobic interiors, or their exceptional hydrophobicity (crambin, ribitol dehydrogenase).

(b) *Tentative classification of helices with the hydrophobic moment plot*

A "hydrophobic moment plot" displays the hydrophobic moment of a helix as a function of its hydrophobicity. In a preliminary study we found that helices of different functional types tend to cluster in different regions of such a plot. Specifically, α-helices from globular, surface and transmembrane proteins tend to

occupy characteristic regions, although the regions do not have clearly defined boundaries.

In Figures 1 and 2, we consider 56 helices from 33 surface proteins and transmembrane proteins from three classes: (1) class I histocompatibility antigens which are probably anchored in the membrane by a single α-helix; (2) membrane bound immunoglobulins and class II histocompatibility antigens (which are dimeric and may have transmembrane M segments that may be dimeric in the membrane (Rogers *et al.*, 1980)); and (3) channel-forming membrane proteins anchored in the membrane by several proteins segments. Each point in Figures 1 and 2 represents a single, 11-residue α-helix. For the transmembrane helices, the 11-residue helix is the fragment of the 21-residue helix, determined above, having the largest hydrophobic moment. For surface proteins, the 11-residue helix is the fragment having the largest hydrophobic moment of the entire sequence. The rationale for plotting 11-residue helices is that their length of about three turns represents the approximate distance over which neighboring helices generally interact. Thus, any tendency to pair larger hydrophobic moments (that is, to shield hydrophilic residues from surrounding lipid) might be better reflected in 11-residue than in 21-residue helices. Helices from globular proteins generally plot in the "globular" region; they are omitted from Figure 1 for the sake of clarity. Also omitted for simplicity are helices from the remaining membrane-related proteins of Table 2.

Helices of the different protein classes show a tendency to plot in different regions of the diagram, as shown by Figure 1. Eight small, surface-seeking

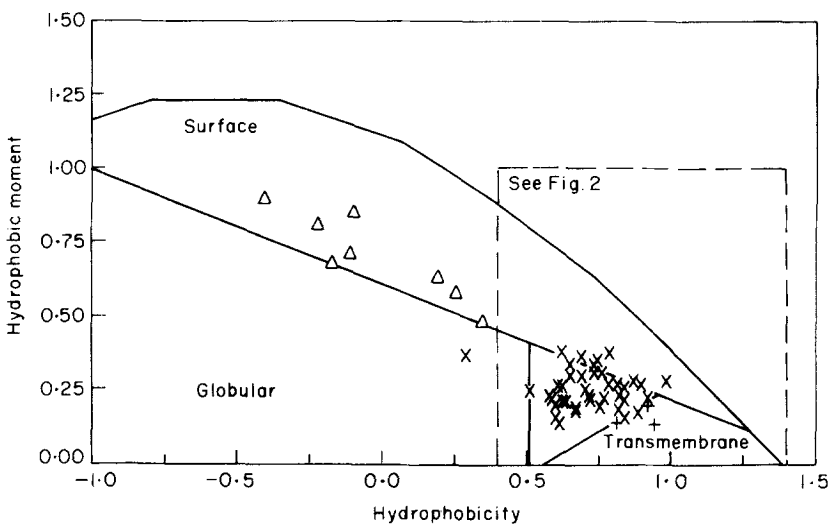

FIG. 1. Hydrophobic moment plot for the protein segments of Table 3. Each point represents an 11-residue α-helix, the ordinate giving the mean hydrophobic moment per residue $\langle \mu_H \rangle$, and the abscissa giving the mean hydrophobicity per residue $\langle H \rangle$. Points labeled Δ are from Table 3; + are from Table 3B1; and × are from Table 3B2 and 3B3. The curve defines the outer limit of possible points, as explained in Methods.

TABLE 3

α-Helices in the hydrophobic moment plots of Figures 1 and 2

Protein	Starting residue	Plot co-ordinates $\langle H \rangle$	Plot co-ordinates $\langle \mu_H \rangle$	*Reference*
A. *Surface region*				
Cecropin A	3 (Lys)	−0·22	0·80	Steiner *et al.* (1981)
Cecropin B	3 (Lys)	−0·41	0·89	
δ-Hemolysin				
S. aureus	16 (Ile)	−0·11	0·70	Fitton *et al.* (1980)
S. aureus (canine strain)	16 (Ile)	−0·17	0·67	
Melittin				
A. mellifera	12 (Gly)	0·25	0·57	Habermann (1972)
A. florea	12 (Gly)	0·34	0·47	Kriel (1973)
Synthetic cytotoxin	12 (Leu)	0·19	0·62	DeGrado *et al.* (1981)
Synthetic cytotoxin	2 (Leu)	−0·10	0·84	DeGrado (personal communication)
B. *Transmembrane region*				
1. *Possible monomers*				
Histocompatibility antigens:				
Mouse class I				
H-2d no. 1	C-43 (Met)	0·92	0·20	Bregegere *et al.* (1981) Moore *et al.* (1982)
H-2d no. 2	C-43 (Val)	0·94	0·13	Kvist *et al.* (1981)
H-2k^b	C-59 (Leu)	0·81	0·14	Coligan *et al.* (1981) Reyes *et al.* (1982)
H-2k^d	C-59 (Leu)	0·81	0·14	Lalanne *et al.* (1983) Kvist *et al.* (1983)
Human class I				
HLA	C-55 (Ile)	0·88	0·18	Malissen *et al.* (1982)
2. *Possible dimers*				
Histocompatibility antigens				
Mouse class II				
A_α	C-29 (Val)	0·87	0·18	Benoist *et al.* (1983)
A_β	C-78 (Leu)	0·70	0·24	Choi *et al.* (1983)
E_α	C-38 (Val)	0·82	0·18	McNicholas *et al.* (1982)
Mouse immunoglobulin heavy chains				
IgA (alpha)	C-41 (Ala)	0·51	0·24	Word *et al.* (1983)
IgD (delta)	C-15 (Leu)	0·65	0·29	Cheng *et al.* (1982)
IgE (epsilon)	C-46 (Phe)	0·72	0·21	Ishida *et al.* (1982)
IgG1 (gamma 1)	C-45 (Phe)	0·75	0·19	Yamawaki-Kataoka *et al.* (1982) Rogers *et al.* (1981) Tyler *et al.* (1982)
IgG2a (gamma 2a)	C-45 (Phe)	0·75	0·19	Yamawaki-Kataoka *et al.* (1982)
IgG2b (gamma 2b)	C-45 (Phe)	0·75	0·19	Yamawaki-Kataoka *et al.* (1982) Rogers *et al.* (1981)
IgG3 (gamma)	C-45 (Phe)	0·75	0·19	Komaromy *et al.* (1983) Lee *et al.* (1982)
IgM (mu)	C-20 (Ile)	0·82	0·26	Rogers *et al.* (1980)
Human class II				
DC1	C-29 (Val)	0·84	0·14 (MON)	Auffray *et al.* (1982)

TABLE 3 *(continued)*

Protein	Starting residue	Plot co-ordinates $\langle H \rangle$	Plot co-ordinates $\langle \mu_H \rangle$	Reference
DR_α	C-29 (Val)	0·92	0·21	Korman *et al.* (1982)
DR_β	C-30 (Leu)	0·78	0·26	Larhammer *et al.* (1982)
3. *Possible channels*				
Acetylcholine receptor				
α-Subunit:				
1	244 (Ile)	0·69	0·29	Noda *et al.* (1982)
2	278 (Thr)	0·87	0·27 (SUR)	
3	304 (Phe)	0·84	0·25	
4	437 (Val)	0·90	0·26 (SUR)	
β-Subunit:				
1	243 (Phe)	0·79	0·37 (SUR)	Noda *et al.* (1983*b*)
2	282 (Ala)	0·72	0·22	
3	313 (Ile)	0·87	0·27 (SUR)	
4	466 (Val)	0·82	0·26	
γ-Subunit:				
1	237 (Leu)	0·75	0·34 (SUR)	Noda *et al.* (1983*a*)
2	271 (Leu)	0·58	0·22	
3	306 (Phe)	0·89	0·17 (MON)	
4	471 (Ile)	0·82	0·22	
δ-Subunit:				
1	247 (Leu)	0·65	0·33	Noda *et al.* (1983*b*)
2	276 (Glu)	0·29	0·36 (GLOB)	
3	313 (Leu)	0·76	0·30 (SUR)	
4	481 (Ile)	0·77	0·22	
Bacteriorhodopsin:				
1	15 (Leu)	0·62	0·25	Khorana *et al.* (1979)
2	42 (Phe)	0·69	0·36 (SUR)	
3	93 (Leu)	0·63	0·20	
4	114 (Ala)	0·60	0·19	
5	136 (Val)	0·67	0·18	
6	183 (Ser)	0·59	0·21	
7	203 (Ile)	0·62	0·37 (SUR)	
Rhodopsin:				
1	49 (Met)	0·67	0·17	Dratz & Hargrave (1983)
2	81 (Val)	0·74	0·30	
3	114 (Gly)	0·61	0·25	
4	162 (Val)	0·62	0·20	
5	210 (Val)	0·99	0·27 (SUR)	
6	258 (Val)	0·84	0·21	
7	286 (Ile)	0·62	0·37 (SUR)	

References are to the sources of the amino acids sequences used in computations. The few segments that plot in a region other than the expected one are indicated by the terms MON (monomeric), SUR (surface) and GLOB (globular) after column 4.

† Sequence numbers preceded by C indicate numbering from the C terminus

peptides which promote cell lysis (Table 3A) all plot at high values of μ_H, and this region is accordingly labeled "surface". The somewhat arbitrary line that divides this region from others has been drawn through points representing melittin from *Apis florea* and δ-hemolysin from *Staphylococcus aureus*. The equation for the line is given by $\langle \mu_H \rangle = (-0{\cdot}392) + (0{\cdot}603)\ \langle H \rangle$.

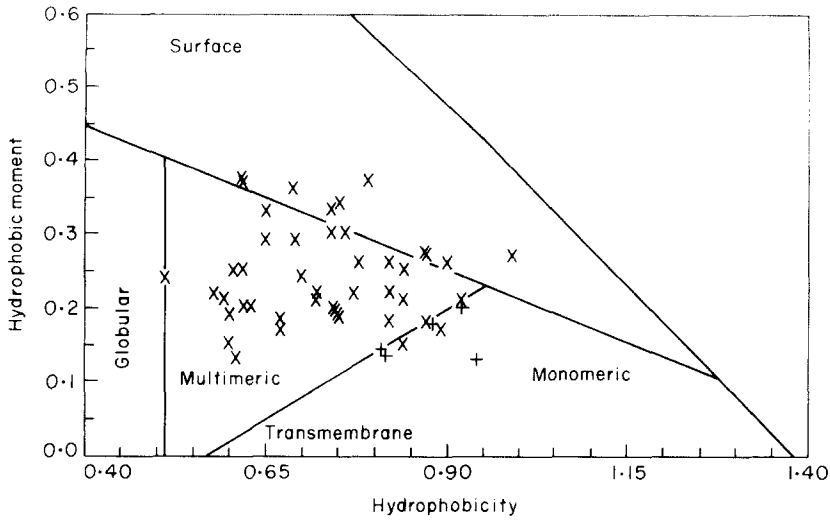

FIG. 2. An enlargement of the boxed region of Fig. 1.

Figure 2 shows the helices from membrane proteins that plot in the region with $\langle H \rangle$ greater than 0·51. (Like the other boundary lines in Figs 1 and 2, this one is arbitrary, selected only because it separates proteins of different functional classes.) Among the 48 helices, there are further patterns. Of seven that plot below the line $\langle \mu_H \rangle = 0{\cdot}600 - 0{\cdot}342 \langle H \rangle$, five represent the class I HLA transmembrane sequences. These sequences are believed to be single-chain membrane anchors, which are not known to make contacts with other protein chains. Accordingly, this region of the plot is termed "monomeric". Above the same line are found the points corresponding to dimeric membrane proteins (such as membrane immunoglobulin chains) and to probable helices in channel-forming proteins such as bacteriorhodopsin, rhodopsin, and the four protein chains of acetylcholine receptor.

The amino acid composition of two of these helical segments illustrates what influence the composition has on their placement on the hydrophobic moment plot. In Figure 1, the leftmost × (least hydrophobic membrane segment) is in helix 2 of the δ subunit of acetylcholine receptor. Among hydrophilic residues, it contains one Glu, one Lys, one Thr and two Ser residues. The rightmost × (most hydrophobic segment) in both Figures 1 and 2 is in helix 5 of rhodopsin. The only hydrophilic residues are His and Pro. However, because they are spaced by four residues they are on the same side of the helix and produce a relatively large hydrophobic moment.

(c) *Predictions for diphtheria toxin and comparisons to other proteins*

Diphtheria toxin exhibits functional characteristics of globular, membrane and surface proteins. It is a single polypeptide chain, soluble in aqueous solution (Collier, 1975), which can bind to planar lipid bilayers, forming transmembrane,

anion-selective channels (Donovan *et al.*, 1981). Moreover, lipid bilayers are disorganized by a cyanogen bromide fragment of diphtheria toxin, as they are by melittin and other surface-seeking proteins (Kayser *et al.*, 1981).

Our algorithm for detecting hydrophobic helices finds five in diphtheria toxin (Table 2C), distinguishing it from globular proteins. At the amino terminus, there is the hydrophobic leader sequence, not present in the mature toxin. In addition, toward the carboxy terminus, in the portion of the protein destined to become the membrane-inserting B fragment, there are four probable transmembrane sequences and one segment of nine consecutive 11-residue windows (centered at residues 361 to 369) that plot in the "surface" region of Figure 1. This segment is contained within the cyanogen bromide fragment that disrupts lipid bilayers. Thus, the present methods reveal segments in the amino acid sequence that may be identified with the membrane-penetrating and surface-binding functions of this unusual protein.

4. Discussion

(a) *Can membrane-associated segments of a protein be recognized from the amino acid sequence?*

This question has been considered by many authors, including general discussions by von Heijne (1981), Steitz *et al.* (1982), Kyte & Doolittle (1982) and Argos *et al.* (1982). There is general agreement that membrane-penetrating segments are more hydrophobic on average than globular segments, but this does not explain all observations. For example, Kyte & Doolittle (1982) noted that a 19-residue segment from a soluble, globular protein (residues 23 to 41 in dogfish lactate dehydrogenase) is more hydrophobic than eight of the nine membrane-associated segments identified by them. If considered to be a transmembrane helix, it would plot at co-ordinates (0·56, 0·16) in the left central portion of the "multimeric" region of Figure 2. Similarly, among seven globular proteins of Table 3C (excluding diphtheria toxin), we find seven that contain at least one (and as many as 4) 21-residue segments having $\langle H \rangle \geq 0{\cdot}42$, the threshold we take for transmembrane candidates (see Methods).

Thus, the hydrophobicity of a protein segment is not sufficient to determine if it is membrane-associated. Another factor that could influence the tendency of a segment to be membrane-associated, is the co-operativity of binding, and this can be incorporated into the model. To do so we suppose that membrane penetration is energetically favorable for *single* segments of polypeptide chains only if they are very hydrophobic, such as the monomer chains of Table 3B. However, less hydrophobic segments can be stable in membranes if associated with other chains, implying some co-operativity of association of helices within the membrane. In this model, dimers or channel-forming segments are stable in the membrane because they enjoy some association with each other. Conversely, the hydrophobic segments from lactate dehydrogenase and ribitol dehydrogenase would not be expected to associate with membranes because they are neither highly hydrophobic nor can they associate with other strongly hydrophobic segments of the same polypeptide.

This explanation of co-operativity accounts well for the membrane-binding capacity of the membrane proteins of Table 2 and for the lack of membrane-binding capacity of the globular controls. Nearly all probable membrane-penetrating segments are identified by assuming that: (1) penetration by a single helix requires a mean hydrophobicity of 0·68 or more; (2) penetration by two or more helices requires either that one of them qualify by virtue of its own high hydrophobicity as in (1) above, or that two of them together have enough hydrophobicity (i.e. have summed $\langle H \rangle$ values of 1·10 or more) to be able to co-operatively assist one another into the membrane. After this "initiation" of membrane binding, other helices may bind if they qualify with $\langle H \rangle \geq 0{\cdot}42$.

Consistent with this idea of co-operative membrane association are two heavy chain transmembrane M segments of IgA and IgM. Eleven-residue fragments in the α and μ heavy chains have $\langle H \rangle$ values of 0·61 and 0·65, respectively, values somewhat low for as penetrating monomers. However, immunoglobulin heavy chains are linked as dimers in immunoglobulin molecules and the summed $\langle H \rangle$ values (1·22 and 1·30, respectively) are well in the range of hydrophobicity required for transmembrane penetration.

A broader view of co-operative association is that all protein–protein, protein–lipid and protein–solvent interactions contribute to the stability of the actual structure, and that structural segments larger than a pair of helices must be considered to determine correctly which segments are membrane bound. However, it appears that the algorithm used here, involving initial co-operativity of two helices, is sufficient to account for nearly all identified membrane-associated proteins.

Table 2A shows three molecules that have been identified by others as membrane-associated proteins, in which our algorithm identifies no membrane-associated segments. The first is respiratory NADH dehydrogenase from *Escherichia coli*. The second is gene product P from the histidine transport operon of *Salmonella typhimurium*, which in preparation appears approximately 20% in the soluble fraction (G. F.-L. Ames, personal communication), and could conceivably be associated with the membrane solely through subunit interactions with the other components of the transport system. The third is matrix porin, from the outer membrane of *E. coli*. Solution studies (Rosenbusch, 1974) suggest little if any α-helix in this protein and some β-sheet. Our failure to detect transmembrane segments in this latter protein could be due to either: (1) its membrane association being related to its oligomerization; or (2) its membrane-associated regions forming a β-structure, such as a β-barrel, that we are unable to detect. The structure of this protein is under study by X-ray crystallography and will presumably illuminate the puzzle (Garavito & Rosenbusch, 1980).

(b) *Can surface-binding peptides be recognized from their amino acid sequences?*

The hydrophobic moment plot seems effective in identifying proteins, including all those in Table 3A, which are known to seek surfaces and to bind to and disrupt cell surfaces. All these proteins are small peptides, having a substantial fraction

(at least 22%) of 11-residue segments falling in the "surface" region (excluding from the total the number that fall in the transmembrane region; Table 2). It is intriguing that a segment of diphtheria toxin also plots in the "surface" region, and it will be interesting to learn if this segment functions in binding or disorganization at the cell surface.

Several other proteins in Table 2 contain several successive 11-residue windows that plot in the "surface" region. These are hisP (6 windows), hisM (8 windows) and phosphorylase (11 windows). The windows in phosphorylase are centered at residues 616 to 626. This segment is known to form a helix exposed to solvent on one side and deeply buried in a hydrophobic region on the other (E. Goldsmith, personal communication).

(c) *Can various classes of membrane association be detected from amino acid sequences?*

The results summarized in Figures 1 and 2 suggest that it may be possible to distinguish sequences that form monomeric membrane association, from those which form multimeric associations, and those that form surface associations. In the small sample considered here, the monomeric anchors are more hydrophobic, and less amphiphilic than transmembrane segments that are likely to be dimeric or are folded into larger membrane helices. This is consistent with the suggestion that bacteriorhodopsin is an "inside-out protein", having the hydrophilic faces of helices inside and hydrophobic faces outside (Engelman & Zaccai, 1980). Indeed, a large amphiphilicity of helices in "channel-forming" proteins might be expected to be a common feature, and the observation of relatively large hydrophobic moments in the putative helices of the channel-forming proteins (Table 3A) is consistent with this idea. Amphiphilic helices in the ion channel of acetylcholine receptor have been proposed by Finer-Moore & Stroud (1984) on the basis of similar methods to ours.

Our suggestion that various classes of protein–membrane association may be detected from amino acid sequences must, of course, be tested by further structural studies of membrane proteins.

We acknowledge with thanks suggestions by Dr R. M. Sweet, Dr D. C. Rees, and Dr R. J. Collier, as well as support from National Institutes of Health grants GM31299, AI13410 and CA12800, and National Science Foundation grant PCM8207520.

REFERENCES

Anderson, S., de Bruijn, M. H. L., Coulson, A. R., Eperon, I. C., Sanger, F. & Young, I. G. (1982). *J. Mol. Biol.* **156**, 683–717.

Argos, P., Rao, J. K. M. & Hargrave, P. A. (1982). *Eur. J. Biochem.* **128**, 565–575.

Auffray, C., Korman, A. J., Roux-Dosseto, M., Bono, R. & Strominger, J. L. (1982). *Proc. Nat. Acad. Sci., U.S.A.* **79**, 6337–6341.

Benoist, C. O., Mathis, D. J., Kanter, M. R., Williams, V. E. II & McDevitt, H. O. (1983). *Proc. Nat. Acad. Sci., U.S.A.* **80**, 534–538.

Bloxham, D. P., Parmelee, D. C., Kumar, S., Wade, R. D., Ericsson, L. H., Neurath, H., Walsh, K. A. & Titani, K. (1981). *Proc. Nat. Acad. Sci., U.S.A.* **78**, 5381–5385.

Boyd, A., Kendall, K. & Simon. M. I. (1983). *Nature (London)*, **301**, 623–626.

Bregegere, F., Abastado, J. P., Kvist, S., Rask, L., Lalanne, J. L., Garoff, H., Kami, B., Wiman, K., Larhammar, D., Peterson, P. A., Gachelin, G., Kourilsky, P. & Dobberstein, B. (1981). *Nature (London)*, **292**, 78–81.

Brewer, H. B. Jr, Lux, S. E., Ronan, R. & John, K. M. (1972). *Proc. Nat. Acad. Sci., U.S.A.* **69**, 1304–1308.

Brewer, H. B. Jr, Shulman, R., Herbert, P., Ronan, R. & Wehrly, K. (1974). *J. Biol. Chem.* **249**, 4975–4984.

Brewer, H. B. Jr, Fairwell, T., LaRue, A., Ronan, R., Houser, A. & Bronzert, T. J. (1978). *Biochem. Biophys. Res. Commun.* **80**, 623–630.

Brown, J. R. & Shockley, P. (1982). In *Lipid–Protein Interactions* (Jost, P. C. & Griffith, O. H., eds), vol. 1, John Wiley and Sons, New York.

Cheng, H.-L., Blattner, F. R., Fitzmaurice, L., Mushinski, J. F. & Tucker, P. W. (1982). *Nature (London)*, **296**, 410–415.

Choi, E., McIntyre, K., Germain, R. N. & Seidman, J. G. (1983). *Science*, **221**, 283–286.

Coligan, J. E., Kindt, T. J., Uehara, H., Martinko, J. & Nathenson, S. G. (1981). *Nature (London)*, **291**, 35–39.

Collier, R. J. (1975). *Bacteriol. Rev.* **39**, 54–85.

Dayhoff, M. O., Hunt, L. T., Barker, W. C., Orcutt, B. C., Yeh, L. S., Chen, K. R., George, D. G., Blomquist, M. C., Fredrickson, J. & Johnson, G. C. (1982). *Protein Sequence Database: April 1982, Tape Format C*, National Biomedical Research Foundation, Washington, D.C.

DeGrado, W. F., Kézdy, F. J. & Kaiser, E. T. (1981). *J. Amer. Chem. Soc.* **103**, 679–681.

Donovan, J. J., Simon, M. I., Draper, R. K. & Montal, M. (1981). *Proc. Nat. Acad. Sci., U.S.A.* **78**, 172–176.

Dratz, E. A. & Hargrave, P. A. (1983). *Trends Biochem. Sci.* **8**, 128–131.

Eisenberg, D., Weiss, R. M. & Terwilliger, T. C. (1982*a*). *Nature (London)*, **299**, 371–374.

Eisenberg, D., Weiss, R. M., Terwilliger, T. C. & Wilcox, W. (1982*b*). *Faraday Symp. Chem. Soc.* **17**, 109–120.

Eisenberg, D., Weiss, R. M. & Terwilliger, T. C. (1984). *Proc. Nat. Acad. Sci., U.S.A.* **81**, 140–144.

Engelman, D. M. & Steitz, T. A. (1981). *Cell*, **23**, 411–422.

Engelman, D. M. & Zaccai, G. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 5894–5898.

Engelman, D. M., Goldman, A. & Steitz, T. A. (1982). *Methods Enzymol.* **88**, 81–88.

Feldman, R. J. (1976). *Atlas of Macromolecular Structure on Microfiche*, Tracor Jitco, Rockville, Md.

Finer-Moore, J. & Stroud, R. M. (1984). *Proc. Nat. Acad. Sci., U.S.A.* **81**, 155–159.

Fitton, J. E., Dell, A. & Shaw, W. V. (1980). *FEBS Letters*, **115**, 209–212.

Foster, D. L., Boublik, M. & Kaback, H. R. (1983). *J. Biol. Chem.* **258**, 31–34.

Garavito, R. M. & Rosenbusch, J. P. (1980). *J. Cell. Biol.* **86**, 327–329.

Gay, N. J. & Walker, J. E. (1981). *Nucl. Acids Res.* **9**, 3919–3926.

Greenfield, L., Bjorn, M. J., Horn, G., Fong, D., Buck, G. A., Collier, R. J. & Kaplan, D. A. (1983). *Proc. Nat. Acad. Sci., U.S.A.* **80**, 6853–6857.

Grundstrom, T. & Jaurin, B. (1982). *Proc. Nat. Acad. Sci., U.S.A.* **79**, 1111–1115.

Habermann, E. (1972). *Science*, **177**, 314–322.

Heinemann, F. S. & Ozols, J. (1982). *J. Biol. Chem.* **257**, 14988–14999.

Henderson, R. (1979). *Soc. Gen. Physiol.* **33**, 3–15.

Henderson, R. & Unwin, P. N. T. (1975). *Nature (London)*, **257**, 28–32.

Hendrickson, W. A. & Teeter, M. M. (1981). *Nature (London)*, **290**, 107–113.

Higgins, C. F., Haag, P. D., Nikaido, K., Ardeshir, F., Garcia, G. & Ames, G. F.-L. (1982). *Nature (London)*, **298**, 723–727.

Ishida, N., Ueda, S., Hayashida, H., Miyata, T. & Honjo, T. (1982). *EMBO J.* **1**, 1117–1123.

Jackson, R. L., Gotto, A. M. Jr, Lux, S. E., John, K. M. & Fleischer, S. (1973). *J. Biol. Chem.* **248**, 8449–8456.

Jackson, R. L., Baker, H. N., Gilliam, E. B. & Gotto, A. M. Jr (1977). *Proc. Nat. Acad. Sci., U.S.A.* **74**, 1942–1945.

Kayser, G., Lambotte, P., Falmagne, P., Capiau, C., Zanen, J. & Ruysschaert, J.-M. (1981). *Biochem. Biophys. Res. Commun.* **99**, 358–363.

Khorana, H. G., Gerber, G. E., Herlihy, W. C., Gray, C. P., Anderegg, R. J., Nihei, K. & Biemann, K. (1979). *Proc. Nat. Acad. Sci., U.S.A.* **76**, 5046–5050.

Komaromy, M., Clayton, L., Rogers, J., Robertson, S., Kettman, J. & Wall, R. (1983). *Nucl. Acids Res.* **11**, 6775–6785.

Korman, A. J., Auffray, C., Schamboeck, A. & Strominger, J. L. (1982). *Proc. Nat. Acad. Sci., U.S.A.* **79**, 6013–6017.

Kriel, G. (1973). *FEBS Letters,* **33**, 241–244.

Kvist, S., Bregegere, F., Rask, L., Cami, B., Garoff, H., Daniel, F., Wiman, K., Larhammar, D., Abastado, J. P., Gachelin, G., Peterson, P. A., Dobberstein, B. & Kourilsky, P. (1981). *Proc. Nat. Acad. Sci., U.S.A.* **78**, 2772–2776.

Kvist, S., Roberts, L. & Dobberstein, B. (1983). *EMBO J.*, **2**, 245–254.

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

Lalanne, J.-L., Delarbre, C., Gachelin, G. & Kaurilsky, P. (1983). *Nucl. Acids Res.* **11**, 1567–1577.

Larhammer, D., Schenning, L., Gustafsson, K., Wiman, K., Claesson, L., Rask, L. & Peterson, P. A. (1982). *Proc. Nat. Acad. Sci., U.S.A.* **79**, 3687–3691.

Lee, J. S., Trowsdale, J., Travers, P. J., Carey, J., Grosveld, F., Jenkins, J. & Bodmer, W. F. (1982). *Nature (London)*, **299**, 750–752.

Malissen, M., Malissen, B. & Jordan, B. R. (1982). *Proc. Nat. Acad. Sci., U.S.A.* **79**, 893–897.

McNicholas, J., Steinmetz, M., Hunkapillar, T., Jones, P. & Hood, L. (1982). *Science*, **218**, 1229–1232.

Moore, K. W., Sher, B. T., Sun, Y. H., Eakle, K. A. & Hood, L. (1982). *Science*, **215**, 679–682.

Muller, K.-D., Salnikow, J. & Vater, J. (1983). *Biochim. Biophys. Acta*, **742**, 78–83.

Noda, M., Takahashi, H., Tanabe, T., Toyosato, M., Furutani, Y., Hirose, T., Asai, M., Inayama, S., Miyata, T. & Numa, S. (1982). *Nature (London)*, **299**, 793–797.

Noda, M., Takahashi, H., Tanabe, T., Toyosato, M., Kikyotani, S., Furutani, Y., Hirose, T., Takashima, H., Inayama, S., Miyata, T. & Numa, S. (1983*a*). *Nature (London)*, **302**, 528–532.

Noda, M., Takahashi, H., Tanabe, T., Toyosato, M., Kikyotani, S., Hirose, T., Asai, M., Takashima, H., Inayama, S., Miyata, T. & Numa, S. (1983*b*). *Nature (London)*, **301**, 251–255.

Reyes, A. A., Schold, M., Itakura, K. & Wallace, R. B. (1982). *Proc. Nat. Acad. Sci., U.S.A.* **79**, 3270–3274.

Rogers, J., Early, P., Carter, C., Calame, K., Bond, M., Hood, L. & Wall, R. (1980). *Cell*, **20**, 303–312.

Rogers, J., Choi, E., Souza, L., Carter, C., Word, C., Kuehl, M., Eisenberg, D. & Wall, R. (1981). *Cell*, **26**, 19–27.

Rosenbusch, J. P. (1974). *J. Biol. Chem.* **249**, 8019–8029.

Segrest, J. P. & Feldman, R. J. (1974). *J. Mol. Biol.* **87**, 853–858.

Shinozaki, K. & Sugiura, M. (1982). *Gene*, **20**, 91–102.

Shulman, R. S., Herbert, P. N., Wehrly, K. & Frederickson, D. S. (1975). *J. Biol. Chem.* **250**, 182–190.

Steiner, H., Hultmark, D., Engstrom, A., Bennich, H. & Boman, H. G. (1981). *Nature (London)*, **292**, 246–248.

Steitz, T. A., Goldman, A. & Engelman, D. M. (1982). *Biophys. J.* **37**, 124–125.

Tyler, B. M., Cowman, A. F., Gerondakis, S. D., Adams, J. M. & Bernard, O. (1982). *Proc. Nat. Acad. Sci., U.S.A.* **79**, 2008–2012.

von Heijne, G. (1981). *Eur. J. Biochem.* **116**, 419–422.

Wickner, W. (1980). *Science,* **210**, 861–868.
Wolfe, P. B., Wickner, W. & Goodman, J. M. (1983). *J. Biol. Chem.* **258**, 12073–12080.
Word, C. J., Mushinski, J. F. & Tucker, P. W. (1983). *EMBO J.* **2**, 887–898.
Yamawaki-Kataoka, Y., Nakai, S., Miyata, T. & Honjo, T. (1982). *Proc. Nat. Acad. Sci., U.S.A.* **79**, 2623–2627.
Young, I. G., Rogers, B. L., Campbell, H. D., Jaworowski, A. & Shaw, D. C. (1981). *Eur. J. Biochem.* **116**, 165–170.

Edited by M. Gellert